The many faces of the chromatin: from genome organization to gene expression

Anil Özdemir

For my parents.

Printed by: Ridderprint BV

The many faces of the chromatin:
from genome organization to gene expression

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus, prof. dr. Th.L.M. Engelen,
volgens besluit van het college van decanen
in het openbaar te verdedigen op woensdag 8 juli 2015
om 16:30 uur precies

door

Anil Özdemir

geboren op 18 oktober 1978
te Zonguldak (Turkije)

Promotor: Prof. dr. ir. H.G. Stunnenberg

Copromotor: Dr. C. Logie

Manuscriptcommissie:

Prof. dr. G.J.C. Veenstra

Prof. dr. P. Verrijzer (Erasmus MC Rotterdam)

Dr. A. Schenck

5-FOA	5-fluoroorotic acid	FDR	false discovery rate
Ac	acetylation	FGF	fibroblast growth factor
AP	anteroposterior	FITC	fluorescein isothiocyanate
ATP	adenosine triphosphate	*fog*	*folded gastrulation*
BAC	bacterial artificial chromosome	GFP	green fluorescent protein
BDGP	Berkeley Drosophila Genome Project	Grk	Gurken
bHLH	basic helix-loop-helix	HAT(s)	histone acetyltransferase(s)
BMP	bone morphogenetic protein	HC	high confidence
bp	base pairs	HDAC(s)	histone deacetylase(s)
brk	*brinker*	*hkb*	*huckebein*
CAF-1	chromatin assembly factor 1	HMT(s)	histone methyltransferase(s)
cdc	cell division cycle	HP1	heterochromatin protein 1
CHD1	chromodomain helicase DNA binding protein 1	*htl*	*heartless*
ChIP	chromatin immunoprecipitation	HU	hydroxyurea
CNS	central nervous system	*ind*	*intermediate neuroblasts defective*
CoREST	(co)repressor for element-1-silencing transcription factor	ISWI	imitation switch
CPT	camptothecin	kb	kilobase pairs
CRM(s)	*cis* -regulatory module(s)	kDa	kilodaltons
da	*daughterless*	l	liter
DIG	digoxigenin	*lacZ*	*bacterial beta-galactosidase*
dl	*dorsal*	LSD1	lysine-specific demethylase 1
DNA	deoxyribonucleic acid	M	molar
Dnmt(s)	DNA methyltransferase(s)	MC	medium confidence
dpp	*decapentaplegic*	me	methylation
DSHB	developmental studies hybridoma bank	*Mef2*	*Myocyte enhancer factor 2*
DV	dorsoventral	MEME	multiple expectation maximum for motif elicitation
Ea	Easter	Mes	mesoderm
ECL	enhanced chemiluminescence	*mirr*	*mirror*
EGF	epidermal growth factor	ml	mililiter
emc	*extra macrochaetae*	MMS	methyl methanesulfonate
EMT	epithelial to mesenchymal transition	mRNA	messenger RNA
eve	*even skipped*	MT	malphigian tubule

Ndl	Nudel
ng	nanogram
NHEJ	non-homologous end joining
nM	nanomolar
NuRD	nucleosome remodeling and deacetylase
ºC	degrees Celsius
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
ph	phosphorylation
PHD	plant homeodomain
phm	*phantom*
PMG	posterior midgut
PNS	peripheral nervous system
PolII	RNA Polymerase II
PWM	position weight matrix
Rel	Relish
rho	*rhomboid*
RNA	ribonucleic acid
SAGA	Spt-Ada-Gcn5 acetyltransferase
SDS	sodium dodecyl sulfate
SELEX	systematic evolution of ligands by exponential enrichment
SET	Su(var)3-9, Enhancer of Zeste, Trithorax
sim	*single-minded*
sna	*snail*
SNF	sucrose non fermenting
Snk	Snake
sog	*short gastrulation*
Spz	Spatzle
srp	*serpent*
Su(H)	*Suppressor of Hairless*
SWI	switch
SWR1	Swi2/Snf2-related 1
SWR-C	Swi2/Snf2-related complex
TBS	tris-buffered saline
TF(s)	transcription factor(s)
TGF-β	transforming growth factor-β
ths	*thisbe*
tin	*tinman*
tld	*tolloid*
twi	*twist*
Ub	ubiquitination
USP(s)	ubiquitin-specific protease(s)
UV	ultraviolet
vn	*vein*
vnd	*ventral nervous system defective*
wntD	*wnt inhibitor of Dorsal*
zen	*zerknullt*
Zfh1	*Zinc finger homeodomain 1*
Zld	Zelda
Zn	zinc
µl	microliter
µM	micromolar

General Introduction

Part 1

1.1 Chromatin

Eukaryotic DNA is packaged into a higher order structure called chromatin which functions as a dynamic scaffold in the regulation of various nuclear processes (Johnson and Dent, 2013). The nucleosome is the fundamental building block of chromatin that consist of approximately 147 base pairs (bp) of genomic DNA wrapped in superhelical turns around a histone octamer that comprises a tetramer of H3 and H4 histones and two H2A–H2B dimers (Luger, 2003; Luger et al., 1997). Arrays of nucleosomes that are linked to one another by linker histone (H1) are organized into higher order structures via formation of chromatin fibers that are compacted further to form highly condensed mitotic chromosomes (Ghirlando and Felsenfeld, 2013; Grigoryev and Woodcock, 2012).

Genomic DNA is the template for essential cellular processes like transcription, replication, recombination and repair, therefore, accessibility of DNA to these nuclear processes is crucial for the survival of organisms (Johnson and Dent, 2013). Eukaryotic cells utilize a number of biological mechanisms to modify, disassemble, reassemble and to remodel the nucleosome structures. Two major classes of protein complexes regulate accessibility of the DNA template and the ability of other proteins/complexes (e.g., those in transcription and replication machinery) to function (Swygert and Peterson, 2014). The first class of enzyme complexes, the ATP-dependent chromatin remodeling complexes such as SWItch/Sucrose Non Fermenting (SWI/SNF) and imitation switch (ISWI) alter nucleosome positioning, either by sliding in *cis-,* or displacing the histone octamer whereby exposing the DNA sequences on the surface (Euskirchen et al., 2012; Glatt et al., 2011; Hargreaves and Crabtree, 2011; Narlikar et al., 2013). The second class of enzyme complexes, the chromatin-modifying complexes, covalently modify histones and other chromatin associated non-histone proteins

by adding or removing chemical moieties (Lalonde et al., 2014; Petty and Pillus, 2013; Suganuma and Workman, 2011). These covalent modifications serve as signals that are recognized by specific *trans*-acting factors that in turn organize chromatin structure or recruit additional factors to allow nuclear processes like transcription to proceed (Becker and Workman, 2013; Bonasio et al., 2010; Johnson and Dent, 2013; Kouzarides, 2007).

1.2 Covalent Modifications and Enzyme Complexes

Histones are subject to various post-translational modifications, including acetylation (Ac), methylation (me), phosphorylation (ph), ubiquitination (Ub), and sumoylation. The majority of these post-translational modifications have been shown to occur within the N-terminal tails of each histone (Figure 1) (Gurard-Levin and Almouzni, 2014; Patel and Wang, 2013; Rothbart and Strahl, 2014).

1.2.1 Histone Acetylation

In general, histone tail acetylation that occurs on lysine residues is associated with transcriptionally active regions within the nucleus. Acetylation results in a charge neutralization; weakening the electrostatic interactions between DNA and histones, thereby facilitating the access of nucleosomal DNA to the transcription machinery (Shahbazian and Grunstein, 2007; Suganuma and Workman, 2011; Swygert and Peterson, 2014). As such, hyperacetylation of histone tails is linked to transcriptionally active chromatin, whereas silent or repressed regions of the genome are hypoacetylated (Grunstein, 1997). Acetylation of lysine residues are carried out by histone acetyltransferases (HATs) (Kouzarides, 2007). HATs, such as the SAGA complex in yeast, are recruited to DNA by physically associating with the RNA polymerase II (Nagy and Tora, 2007). Acetylated lysine residues are recognized by a protein domain called bromodomain (Sanchez et al., 2014). Bromodomain containing ATP dependent chromatin remodeling complexes, such as SWI/SNF, bind to acetylated histone tails (or rather recruited); leading to a further nucleosome displacement, thereby keeping the chromatin in an open state (Chandy et al., 2006; Lee and Workman, 2007; Petty and Pillus, 2013; Sanchez

et al., 2014). Conversely, histone deacetylases (HDACs) remove acetyl groups, leading to a more compact, or a less accessible chromatin (Brunmeir et al., 2009; de Ruijter et al., 2003; Gregoretti et al., 2004; Moser et al., 2014; Shahbazian and Grunstein, 2007). For example, the *Drosophila* NuRD complex that contains two histone deactylase subunits is recruited to the homeotic (HOX) gene locus by the transcription factor Hunchback to repress transcription (Kehle et al., 1998).

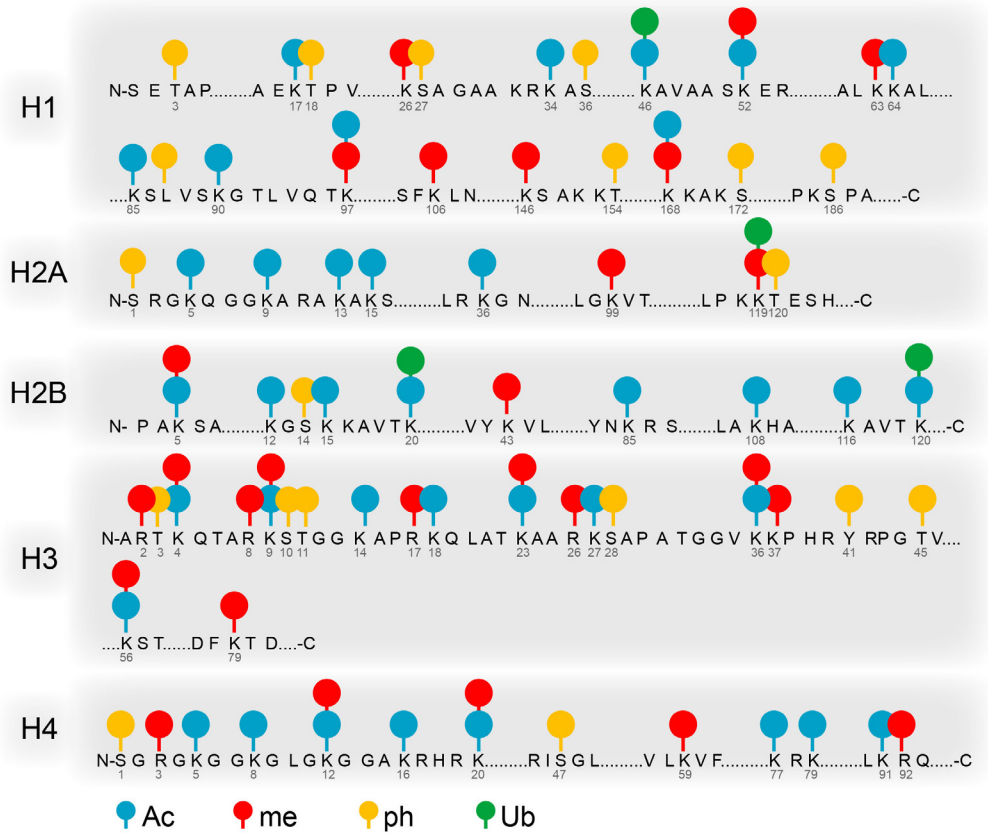


Figure 1: Post-translational modifications on histones.

Acetylation of histone H3 lysine 56 (H3-K56), a core domain residue located near the entry-exit point of the DNA on nucleosome, was first reported in yeast (Masumoto et al., 2005; Ozdemir et al., 2005; Tessarz and Kouzarides, 2014; Xu et al., 2005). Cells lacking H3-K56 acetylation are sensitive to DNA

damaging agents, suggesting that in budding yeast this modification is important for genome integrity (Hyland et al., 2005; Masumoto et al., 2005; Ozdemir et al., 2005; Xu et al., 2005). H3-K56 acetylation occurs on the newly synthesized histone H3 before its deposition into chromatin during S phase, and is removed when cells enter the G2 phase of the cell cycle (Celic et al., 2006; Maas et al., 2006; Masumoto et al., 2005; Recht et al., 2006).

In budding yeast, H3-K56 has been shown to be acetylated by the recently discovered acetyltransferase Rtt109p and the histone chaperone Asf1 has been shown to assist this enzymatic activity (Driscoll et al., 2007; Han et al., 2007; Recht et al., 2006; Schneider et al., 2006; Tsubota et al., 2007; Xhemalce et al., 2007). NAD^+-dependent deacetylases Hst3p and Hst4p have been shown to be the K56 deacetylases (Celic et al., 2006; Maas et al., 2006). Interestingly, cells lacking both Hst3p and Hst4p are sensitive to DNA damaging agents, suggesting hyperacetylation as well as hypoacetylation of H3-K56 is toxic for cells and therefore H3-K56Ac levels must carefully be regulated (Celic et al., 2006; Maas et al., 2006).

Although acetylation of H3-K56 was initially thought to be absent in higher eukaryotes, recent reports have shown that H3-K56 is also acetylated in mammalian cells (Das et al., 2009; Yuan et al., 2009), and more recently in embryonic stem cells (Tan et al., 2013). Human CBP and p300 have been shown to be the H3-K56 acetylases and hGCN5 has been shown to act as an H3-K56 acetylase *in vitro* and is required for H3-K56 acetylation *in vivo*. Whereas SIRT1 and SIRT2 are the human H3-K56 deacetylases (Das et al., 2009; Tjeertes et al., 2009).

H2A.Z is a histone variant that marks nucleosomes at promoters of protein coding genes, chromatin boundary elements, replication origins and centromeres The yeast SWR-C chromatin remodeling enzyme regulates H2A.Z deposition on to chromatin (Gerhold and Gasser, 2014; Krogan et al., 2003a; Mizuguchi et al., 2004; Raisner et al., 2005). It has been recently shown that acetylation of H3-K56 alters substrate specificity of SWR-C dimer-exchange reaction, leading to the removal of H2A.Z from the nucleosomal product (Watanabe et al., 2013). In

the same study, it was also shown that H3-K56 acetylation enhances the ability of the INO80 chromatin remodeling complex to evict H2A.Z/H2B dimers from nucleosomal substrates (Gerhold and Gasser, 2014; Watanabe et al., 2013).

1.2.2 Histone Methylation

Histone methylation occurs on lysine or arginine residues. Unlike acetylation of histone tails, depending on the residue that is modified a methylation mark can be activating or repressive (Black et al., 2012). For example, repetitive DNA regions that are packed in repressed heterochromatin domains are enriched in H3-K9 di- and tri-methylated histones, whereas many actively transcribed genes are associated with di- and tri-methylated H3-K4, K36, and K79 histone marks (Bannister et al., 2002; Bannister et al., 2005; Krogan et al., 2003b; Lalonde et al., 2014; Morris et al., 2005; Ng et al., 2003; Schotta et al., 2002; Schubeler et al., 2004).

Histone tails can be mono-, di- or tri- methylated on lysine residues and mono- or di- methylated on arginine residues by three classes of histone methyltransferases (HMTs) (Del Rizzo and Trievel, 2014). Arginine residues are mono- or di- methylated either symmetrically or asymmetrically by the PRMT family of HMTs (Wysocka et al., 2006). The Su(var)3-9, Enhancer of Zeste, Trithorax (SET) domain HMTs are responsible for the catalysis of all other known histone lysine methylations, except H3-K79 mono-metylation; the DOT1 family of non-SET domain containing family of HMTs are responsible for this particular modification (Del Rizzo and Trievel, 2011; Ng et al., 2002; Qian and Zhou, 2006; van Leeuwen et al., 2002).

Unlike an acetyl group, a methyl group is relatively small and its addition to a lysine or an arginine does not neutralize their charges, although it increases hydrophobicity. Therefore, it is unlikely that a methylation mark alone significantly affects the nucleosome architecture. Evidence suggests that methylated histones recruit proteins that contain specific methyl binding domains that interact with differentially methylated lysine residues to regulate chromatin structure. Thus, many studies have long been focused on identifying and characterizing such

domains and protein complexes (Zentner and Henikoff, 2013). At least four protein domains that have been identified to specifically bind to methylated lysine residues; the plant homeodomain (PHD), chromodomains, Tudor domains, and WD40 repeats (Adams-Cioaba and Min, 2009; Eissenberg, 2012; Gayatri and Bedford, 2014; Migliori et al., 2012; Musselman and Kutateladze, 2011; Vermeulen et al., 2010). For instance, heterochromatin protein 1 (HP1) contains a chromodomain that allows it to specifically recognize methylated lysine 9 of histone H3 (H3-K9me), a mark of repressive chromatin; whereas the budding yeast chromodomain helicase DNA-binding protein 1 (CHD1) recognizes the activating histone mark methylated H3-K4 to enhance transcription (Bannister and Kouzarides, 2005; Jacobs and Khorasanizadeh, 2002; Nielsen et al., 2002; Pray-Grant et al., 2005).

Lysine specific demethylase 1 (LSD1) was the first lysine demethylase identified (Rudolph et al., 2013; Shi et al., 2004). LSD1 type of enzymes have been shown to remove di- or mono- methylated lysines, whereas recently identified another class of demethylases, the Jumonji-class of enzymes can remove all three types of methylation marks (Del Rizzo and Trievel, 2014; Huang et al., 2006; Klose et al., 2006; Shmakova et al., 2014; Tsukada et al., 2006; Yamane et al., 2006). Lysine demethylases have been shown to exist in protein complexes that include other chromatin modifying enzymes. For example, LSD1 is a member of both the CoREST-HDAC and Mi-2/NuRD protein complexes (Lee et al., 2006; Mosammaparast and Shi, 2010; Wang et al., 2009). In both cases, demethylation of the H3-K4me2 by LSD1 complements histone deacetylation.

1.2.3 Histone Phosphorylation

Histone phosphorylation is often associated with transcriptional regulation, response to DNA-damage and mitotic checkpoint pathways (Rossetto et al., 2012). Phosphorylation of H2A variant H2A.X at S139 (γ-H2A.X) (H2A-S129ph in yeast) is one of the best studied histone modifications in response to DNA double strand breaks (DSBs) (Rogakou et al., 1998; Rossetto et al., 2012; Shroff et al., 2004; Stiff et al., 2004). Like most other histone phosphorylation

events (or protein phosphorylation), γ-H2A.X is induced by an intracellular signaling pathway, and is believed to be necessary for the recruitment of the chromatin remodeling complexes, such as INO80 and SWR1 (Downs et al., 2004; Gerhold and Gasser, 2014; Johnson and Dent, 2013; Morrison et al., 2004; van Attikum et al., 2004).

Although most histone proteins are phosphorylated during mitosis, the role/significance of this particular histone modification is poorly understood. Earlier genetic experiments in *Tetrahymena thermophile* have suggested that H3-S10ph is associated with altered chromosome condensation (Wei et al., 1999). However, yeast mutants lacking Ser10 (Ser10Ala) does not show any growth defect and are able to progress through the cell cycle normally (Hsu et al., 2000). Thus, in *S. cerevisiae*, a relationship between H3-S10ph and chromosome dynamics has not been observed.

1.2.4 Histone Ubiquitination and Sumoylation

In addition to being modified by small chemical moieties, histones are also subject to much larger covalent modifications such as ubiquitination and sumoylation (Cubenas-Potts and Matunis, 2013; Fuchs and Oren, 2014). Ubiquitin is a 76 amino acid protein. Histones have shown to be mono- or poly-ubiquitinated (poly-Ub) on lysine residues (Lee et al., 2007; Robzyk et al., 2000; Shukla and Bhaumik, 2007; Weake and Workman, 2008). In higher eukaryotes, histone H2A and H2B are the primary targets of histone ubiquitination (Fuchs and Oren, 2014; Jason et al., 2002). Ubiquitination of histone H2A and H2B has been generally linked to processes like DNA repair, response to DNA damage, protein degradation pathways, gene activation, and silencing (Belle and Nijnik, 2014; Fang et al., 2004; Giannattasio et al., 2005; Huen et al., 2007; Joo et al., 2007; Kalb et al., 2014; Sun and Allis, 2002; van der Knaap et al., 2005; Wozniak and Strahl, 2014). Sequential action of E1, E2, and E3 enzymes has been shown to mediate histone ubiquitination, whereas deubiquintination is catalyzed by a class of thiol proteases known as ubiquitin-specific proteases (USPs) (Nijman et al., 2005; Weake and Workman, 2008).

Small ubiquitin-related modifier (SUMO) is a member of the small ubiquitin like protein family. Although SUMO shares very little sequence identity with ubiquitin (around 18%) and is slightly larger in size (12 and 9 kDa, respectively), they have nearly identical structural fold. Like ubiquitination, sequential action of E1, E2, and E3 enzymes has been shown to mediate attachment of SUMO to other proteins (Nathan et al., 2003). All four core histones have shown to be sumoylated (Cubenas-Potts and Matunis, 2013; Nathan et al., 2006; Shiio and Eisenman, 2003). Evidence suggests that sumoylation of the histone H4 mediates gene silencing through recruitment of histone deacetylase and heterochromatin protein 1 (HP1) (Nathan et al., 2006; Shiio and Eisenman, 2003).

1.3 DNA Methylation

In addition to enzymatic activities that modify chromatin associated histones and non-histone proteins, higher eukaryotic DNA can also be modified at CpG dinucleotides by methylating/demethylating enzymes (Ferguson-Smith and Greally, 2007; Law and Jacobsen, 2010). This reaction is catalyzed by a group of enzymes, the DNA methyltransferases (Dnmts) (Cedar and Bergman, 2009; Kar et al., 2012; Law and Jacobsen, 2010; Robertson et al., 2000; Schaefer and Lyko, 2010). Similar to histone modifications, DNA methylation marks *cis*-regulatory regions, transposable elements, and pericentromeric repeats to alter chromatin compaction and DNA accessibility, and is important for many developmental processes like gene silencing, genomic imprinting and X-chromosome inactivation (Ferguson-Smith, 2011; Gopalakrishnan et al., 2009; Hellman and Chess, 2007; Jin et al., 2008; Macfarlan et al., 2012; Rose and Klose, 2014; Stadler et al., 2011). It has been shown that methylated DNA interferes with the binding of transcription factors to target sites (Birke et al., 2002; Prendergast and Ziff, 1991). In addition, members of the methyl binding proteins have been shown to recruit HDACs and other chromatin remodelers to further promote repressive chromatin environment (Cedar and Bergman, 2009; Nan et al., 1998; Robertson et al., 2000; Rose and Klose, 2014; Rountree et al., 2000; Wade et al., 1999).

Recently it was shown that in pluripotent cells and differentiated mammalian

cell types including human skeletal muscle and brain, non-CpG methylation is associated with gene bodies and it correlates with active transcription (Barrès et al., 2009; Guo et al., 2014; Lister et al., 2009; Ramsahoye et al., 2000; Yan et al., 2011). Although the role of this particular modification remains unclear, it has been implicated in post-transcriptional RNA splicing (Lister et al., 2009).

Part 2

2.1 The Life Cycle of the Fruit Fly *Drosophila melanogaster*

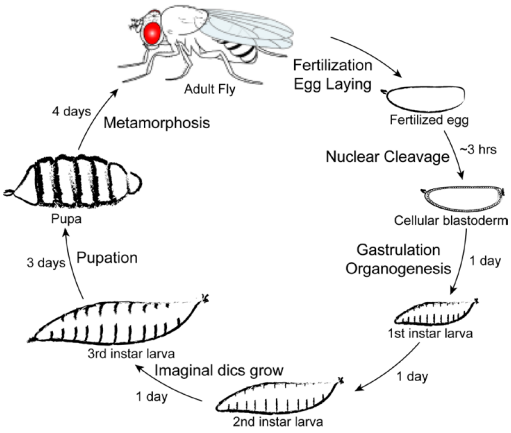


Figure 2 : The life cycle of *Drosophila*.

The life cycle of *Drosophila* occurs over a span of 9-10 days (Figure 2). Embryonic development starts immediately after fertilization of the egg and it takes about 24 hours. Hatching of the embryo gives rise to the larva which feeds, grows and passes through three developmental stages, called instars. During this time molting occurs and the head, mouth, cuticle, spiracles and hooks are shed. At the end of the third instar stage, the pupa is formed. During pupation, an extensive remodeling of the body takes place and the metamorphosis of the fly finally completes about 9 days after fertilization when the adult fruit fly emerges.

2.2 Early Development of the *Drosophila* Embryo

Unlike in higher vertebrates, the *Drosophila* embryo is a syncytium during the first few hours of the development; the nuclei divide and migrate in a common cytoplasm without cell division (Figure 3). During the preblastoderm stage (mitotic cycles 1-9), the nuclei divide rapidly and synchronously until they migrate towards the perimeter of the egg to form a monolayer structure, the syncytial blastoderm. During syncytial blastoderm stage (mitotic cycles 10-

13), zygotic transcription starts, and the rate of the nuclear divisions slows down dramatically. Around 2 hours after fertilization, cellular blastoderm stage begins where the plasma membrane starts to grow inward from the egg surface to enclose each nucleus, eventually creating a single layer of cells around the egg yolk (Figure 3). By the time the cellular blastoderm formation is complete, the *Drosophila* embryo consist of approximately 6000 cells.

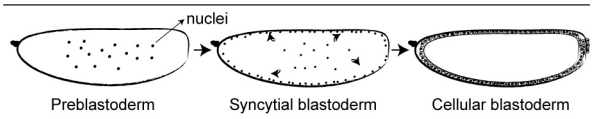


Figure 3: Early *Drosophila* development.

Gastrulation, the formation of germ layers and segregation of the presumptive mesoderm, endoderm, and ectoderm, begins immediately after the cellularization is complete. Through coordinated cell movements, prospective mesoderm cells fold inward to form the ventral furrow which eventually forms a layer of flat mesodermal tissue surrounded by a layer of ectoderm outside.

However, differentiation of tissues begins long before gastrulation starts. Early *Drosophila* embryo can be divided into four regions along the dorsoventral (DV) axis. The ventral domain gives rise to mesodermal tissues (Figure 4). Neurogenic ectoderm gives rise to the nervous system and the ventral epidermis that are formed above the presumptive mesoderm. Finally, dorsal ectoderm which forms the dorsal epidermis and amnioserosa make up the dorsal most regions of the embryo respectively.

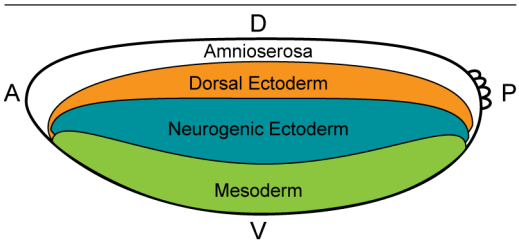


Figure 4: Illustration of an early embryo fate map.

During early *Drosophila* development, differentiation of tissues is established by morphogen gradients. A total of four egg-polarity gene systems help specify the two main axes of the *Drosophila* embryo: the DV axis and the anteroposterior axis (AP) (Figure 5). Each of these systems is responsible for the formation of a particular body part. Along the AP axis, head and the thorax are formed by an anterior system (Frohnhofer et al., 1986), the posterior system forms the abdomen (Lehmann and Nusslein-Volhard, 1986; Nusslein-Volhard et al., 1987), and the terminal system at both ends of the embryo forms the terminal structures (Klingler et al., 1988; Schupbach and Wieschaus, 1986). Along the DV axis a single regulatory system, the Dorsal (Dl) gene regulatory network establishes the polarity (Anderson et al., 1985).

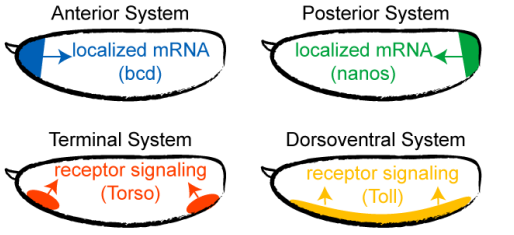


Figure 5: Organization of the four egg-polarity systems.

2.3 Toll Signaling and Formation of DV Patterning

The Toll signaling pathway was originally identified in a series of genetic screens developed in *Drosophila.* Nüsslein-Volhard and Wieschaus who performed the initial screens, identified a number of genes that controls early segmentation (Nusslein-Volhard and Wieschaus, 1980). Together with Ed Lewis, this approach earned them the Nobel Prize in medicine in 1995. Subsequently, analysis of the genetic interactions led to the discovery of 15 genes, called the dorsal group of genes (e.g., *Toll*, *cactus*, NF-κB homolog *dorsal*), as components of an important signaling pathway required for DV patterning of the *Drosophila* embryo (Belvin and Anderson, 1996). Further studies demonstrated that in *Drosophila*, the Toll pathway is involved in both developmental processes as well as immunity (Belvin and Anderson, 1996; Halfon et al., 1995; Lemaitre et al., 1996; Qiu et al., 1998).

The DV polarity is first established in the egg chamber during oogenesis. The EGF ligand Gurken (Grk) - that is associated with the dorsally located oocyte nucleus - signals to nearby follicle cells to repress *Pipe* expression, a gene encoding a putative 2-O sulfotransferase (Figure 6A) (Morgan and Mahowald, 1996; Peri et al., 2002; Sen et al., 1998).

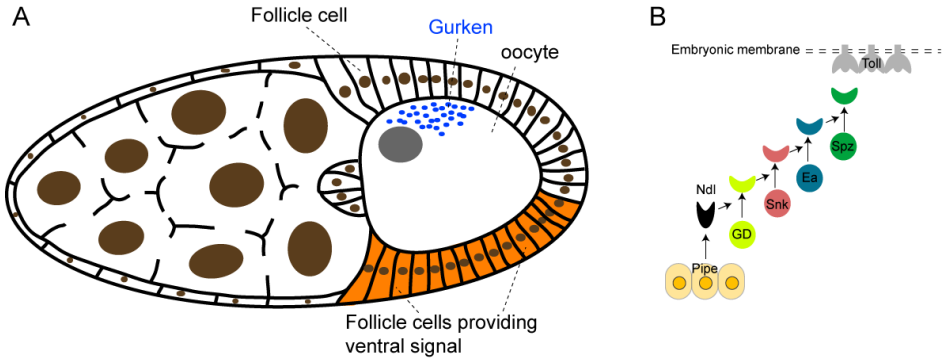


Figure 6: Serine protease activity in the perivitelline space of the *Drosophila* egg

Hence, localized Pipe activity in the ventral cells activates the protease Nudel (Ndl) which is secreted to the perivitelline space, the fluid between the follicle cells and the oocyte (Hong and Hashimoto, 1995). Ndl then initiates a serine protease pathway -involving Gastrulation defective (GD), Snake (Snk), and Easter (Ea)- that ultimately leads to the proteolytic cleavage and the activation of Spatzle ligand (Spz) and the Toll receptor (Figure 6B) (Chasan and Anderson, 1989; Cho et al., 2010; DeLotto and DeLotto, 1998; Jang et al., 2006; Schneider et al., 1994; Stein and Nusslein-Volhard, 1992; Stein et al., 1991). Activated Toll receptor then triggers an intracellular signaling pathway that facilitate degradation of Cactus, a cytoplasmic tethering protein, thereby releasing Dl transcription factor from cytoplasmic retention (Edwards et al., 1997; Geisler et al., 1992; Hecht and Anderson, 1993; Kidd, 1992; Reach et al., 1996; Roth et al., 1991; Towb et al., 2001; Towb et al., 1998). This leads to the nuclear import of Dl at higher levels in the ventral-most cells (due to higher Toll activity), and in gradually lower levels in lateral and dorsal cells; thus creating a ventral to dorsal nuclear gradient (Drier et

al., 1999; Roth et al., 1989; Rushlow et al., 1989).

2.4 The Dorsal Network

dl encodes a sequence specific transcription factor that belongs to the Rel family of transcription factors. Dl is present in a nuclear-cytoplasmic gradient along the DV axis with higher levels of the protein present in ventral regions and lower levels present when progressing more dorsally (Moussian and Roth, 2005; Rushlow and Shvartsman, 2012). The amount of Dl present within nuclei influences levels of gene expression, as do the affinity/number of binding sites within target *cis*-regulatory modules (CRMs) and cooperative interactions with other transcription factors. High levels of Dl in the ventral regions of the embryo activates *twist* (*twi*), *snail* (*sna*), and the fibroblast growth factor receptor (FGFR) *heartless* (*htl*) that are required for the differentiation of the mesoderm (Figure 7) (Ip et al., 1991; Jiang et al., 1991; Pan et al., 1991; Simpson, 1983; Stathopoulos et al., 2004; Thisse et al., 1987). Intermediate levels of Dl activate genes like *rhomboid* (*rho*), *ventral neuroblasts defective* (*vnd*) that are required for the specification of neurogenic ectoderm (Bier et al., 1990; Ip et al., 1992a; Jimenez et al., 1995). Lowest levels of the Dl nuclear gradient activates genes like *short-gastrulation* (*sog*), fibroblast growth factor (FGF) ligand *thisbe* (*ths*) throughout the dorsal ectoderm, and dorsal mesoderm (Francois et al., 1994; Markstein et al., 2002; Stathopoulos et al., 2004; Stathopoulos et al., 2002).

Transcriptional responses of Dl activation also depend on other factors. The transcription factors Daughterless (Da), Grainyhead, STAT92E, Suppressor of Hairless [Su(H)], Twi, and Zelda (Zld) have all been shown to play accessory roles in the activation of gene expression along the DV axis (Garcia and Stathopoulos, 2011; Jiang and Levine, 1993; Liberman and Stathopoulos, 2009; Morel and Schweisguth, 2000). Cooperative interactions between these (and possibly other) factors influence expression along the DV axis (Reeves and Stathopoulos, 2009). For example, Twi is also present in a nuclear gradient, but compared to the Dl gradient, it exhibits a steeper decrease in ventrolateral domains of the embryo. Together these factors are thought to regulate expression of target genes in ventral

and ventrolateral regions of the embryo (e.g., *sna* and *rho* respectively) (Ip et al., 1992b; Jiang and Levine, 1993; Markstein et al., 2004; Zinzen et al., 2006). Whereas in dorsolateral regions of the embryo, cooperative interactions between Dl and Zld help to extend gene expression boundaries further dorsally (Liberman and Stathopoulos, 2009).

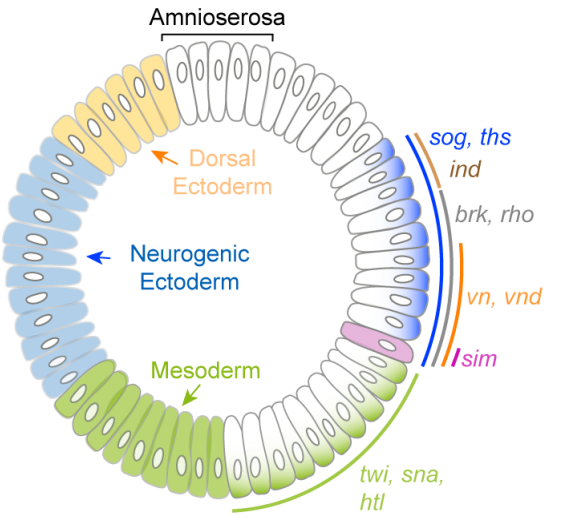


Figure 7: Illustration of early embryonic fate map, and Dl target genes.

2.5 Twi and Sna: Regulators of Mesoderm Differentiation

One of the earliest genes activated by Dl is *twi*. It encodes a basic helix-loop-helix (bHLH) family of transcription factor that is implicated in mesoderm differentiation (Baylies and Bate, 1996; Harfe et al., 1998; Reuter and Leptin, 1994). As well as being a master regulator required for mesoderm specification, it has been shown to convert non-mesodermal cells into mesodermal fate (Baylies and Bate, 1996). Twist has been shown to recognize a core DNA consensus, CANNTG, called an E-box (Massari and Murre, 2000). Twi and Dl cooperatively activate *sna* which encodes a transcription factor, part of a conserved Snail family of zinc finger proteins required for the specification of mesodermal cell fate and invagination of presumptive mesoderm during gastrulation (Leptin and

Grunewald, 1990).

Although both are regulated by Dl, expression domains of *twi* and *sna* show differences; *twi* is expressed in a graded fashion, highest levels seen in the ventral most cells. Its expression also expands beyond the presumptive mesoderm (i.e., *sna* expression border), overlapping with *single-minded* (*sim*) expression in mesectoderm and *rho* in neurogenic ectoderm (Figure 7). *sna* expression on the other hand is robust (i.e., uniform throughout the presumptive mesoderm). *sna* expression border is also sharp, precisely defining the border of mesoderm as it represses expression of mesectodermal and neuroectodermal genes within its expression domain (e.g., *sim*, *rho*, *vnd*, and *sog*) (Figure 7) (Leptin, 1991).

The Dorsal-Twist-Snail gene network has been extensively studied in *Drosophila* to understand how polarity is established by gene regulatory networks (Stathopoulos and Levine, 2005). In the presumptive mesoderm, Twi and Sna regulate expression of several genes that are required for proper mesoderm differentiation and gastrulation. *Myocyte enhancer factor 2* (*Mef2*) (Lilly et al., 1994), *tinman* (*tin*) (Bodmer et al., 1990; Yin et al., 1997) , *htl* (Shishido et al., 1993), *folded gastrulation* (*fog*) (Costa et al., 1994), *Zinc finger homeodomain 1* (*zfh1*) (Casal and Leptin, 1996) and *serpent* (*srp*) (Hemavathy et al., 1997) levels are either substantially reduced or completely absent in *sna* or *twi* deficient embryos. Mef2 is a transcription factor that plays a major role in differentiation of all three muscle types (Cripps et al., 1998). Recent studies have shown Mef2 binds to a large number of enhancers, including ones that regulate transcription factors and differentiation factors, as well as its own (Cripps et al., 1998; Sandmann et al., 2006). The homeodomain transcription factor Tin is necessary for the specification of the heart, the visceral muscle, and a subset of the somatic muscles later during development (Yin and Frasch, 1998; Yin et al., 1997). Zfh1 helps maintain the mesodermal cell fate (Casal and Leptin, 1996). Srp, Fog, and a recently identified transmembrane protein T48 have been shown to induce cell shape changes that are required for the formation of the ventral furrow and invagination (Costa et al., 1994; Dawes-Hoang et al., 2005; Hemavathy et al., 1997; Kolsch et al., 2007).

2.5.1 Twi and Sna: Beyond Regulation of *Drosophila* Morphogenesis

Epithelial cells line cavities and surfaces of other tissues throughout the body to form a protective barrier. The transmembrane protein E-cadherin is required to maintain the tight contact and polarity between epithelial cells. The epithelial to mesenchymal transition (EMT) is a process whereby epithelial cells lose their epithelial features and acquire fibroblast-like characteristics and morphology. As such, EMT is an important step during malignant tumor progression. During development, EMT is induced by several different mechanisms including receptor tyrosine kinases, the transforming growth factor-β (TGF-β)and bone morphogenetic protein (BMP) pathway, and Wnt signaling (Barrallo-Gimeno and Nieto, 2005; Nieto, 2002; Thiery and Sleeman, 2006). In humans, SNAIL1 is the master EMT inducer that initiates repression of CDH1 (the gene encoding E-cadherin), and upregulation of proteins involved in cell motility and extracellular matrix remodeling (Batlle et al., 2000; Cano et al., 2000). TWIST1 has been shown to regulate cell invasion and migration during later stages of EMT (Yang et al., 2004). SNAIL1 has been shown to recruit corepressors like the Smad2/4 complex, and SIN3A as well as chromatin modifying factors such as LSD1, HDAC1, and HDAC2 to CDH1 promoter elements (Hemavathy et al., 2000). Interplay between SNAIL1 and NF-κB (human homologue of the *Drosophila* Dl protein) has also been shown to be required for activation of target genes that regulate downstream events during EMT, suggesting that SNAIL1 cofactors present in the nucleus might regulate its function to act as a repressor or an activator during mesenchymal differentiation (Peinado et al., 2007; Stanisavljevic et al., 2011).

There is growing evidence from *Drosophila* and other systems to indicate that Sna and Twi (and their homologs) might play a wider range of roles than just being master regulators of muscle differentiation and gastrulation. These include cell proliferation, survival, neuronal differentiation, maturation of neural stem cells, and macrophage differentiation (Barrallo-Gimeno and Nieto, 2005; Boutet et al., 2007; Perez-Losada et al., 2003; Saeed et al., 2014; Yang et al., 2010).

2.6 *cis*-Regulatory Logic and the DV Patterning

Combinatorial regulation is one of the major challenges in transcriptional regulation (Weingarten-Gabbay and Segal, 2014). Overlapping actions of activators and repressors determine many complex transcriptional outputs (Levine and Tjian, 2003; Stathopoulos and Levine, 2005). *Drosophila* gastrulation is an excellent model system to dissect the mechanisms of combinatorial regulation (Rushlow and Shvartsman, 2012; Spitz and Furlong, 2012). For example, synergistic interaction between Dl and Twi is required to regulate gene expression in more lateral regions of the embryo where neither factor alone is able to maintain gene expression independently (i.e., *rho*, *vn*, and *vnd* expression domains, Figure 6) (Reeves and Stathopoulos, 2009; Rushlow and Shvartsman, 2012).

Transcription factors are DNA-binding proteins that recruit enzymatic activities to control either chromatin structure (modify DNA/histones) or chromatin environment (recruit coactivators/corepressors), ultimately affecting expression of their target genes (Fuda et al., 2009; Lelli et al., 2012; Spitz and Furlong, 2012). Since the *cis*-regulatory elements (promoter/enhancer sequence) of a gene are the binding sites for transcription factors, identifying/analyzing *in vivo* DNA-binding sequences of a transcription factor is essential for understanding of cellular processes (Buecker and Wysocka, 2012; Levo and Segal, 2014; Ong and Corces, 2011). Most of our knowledge about the Dorsal-Twist-Snail gene network logic comes from classical genetic screens (Bier et al., 1990; Ferguson and Anderson, 1992a; Ferguson and Anderson, 1992b; Jazwinska et al., 1999; Jimenez et al., 1995; Klambt et al., 1989; Kosman et al., 1991; Leptin and Grunewald, 1990; Nambu et al., 1990; Rushlow and Levine, 1990; Rushlow and Shvartsman, 2012; Vaessin et al., 1990). Early experiments used laborious methods to dissect the regulatory logic the Dorsal network operated on, and succeeded in identification of only a handful of *cis*-regulatory regions. It was only recently that by the use whole-genome technologies (i.e., chromatin immunoprecipitation followed by chip) thousands of new putative *cis*-regulatory regions were identified (Sandmann et al., 2007; Zeitlinger et al., 2007). Among these however, only a few have been experimentally tested to function as enhancers.

The main difficulty in identifying *cis*-regulatory elements for the Dorsal-Twist-Snail regulatory network (or any gene regulatory network) is that these sequences are generally short, and degenerate. For instance, a transcription factor binding site could support binding of several members of a family of transcription factors, or closely related factors (e.g. *in vitro* evidence suggest that both Twist and Snail share the same binding site, so are other members of bHLH family of transcription factors), or that a binding site may never be occupied if a sequence element recognized by a cofactor is not occupied under certain conditions (Kellis et al., 2014; Levo and Segal, 2014; Shlyueva et al., 2014).

How the interactions between these transcriptional regulators affect their binding to diverse target loci is also a major question. Current evidence suggests that there are competitive and cooperative interactions for recruitment to some target genes (Plank and Dean, 2014; Stathopoulos and Levine, 2005). It is not known however how the effects of these regulatory actions (cooperation or competition) relate to transcriptional control. Furthermore, whether combinatorial regulation is a general, genome-wide mechanism of Dorsal-Twist-Snail gene network for regulating *Drosophila* gastrulation is also not clear.

Thesis Outline

The aim of this thesis was to study how histone post-translational modifications and transcription factors control higher order structure of chromatin to regulate transcriptional or DNA damage checkpoint responses in yeast and in fruit-fly. Chapter 2 describes characterization of histone H3 lysine 56 acetylation as a novel core domain histone modification in yeast. Chapter 3 provides a detailed analysis of acetylation, deactylation and cell-cycle regulation of histone H3 lysine 56. Chapter 4 describes high resolution mapping of transcription factor Twist to DNA in early *Drosophila* embryos. Chapter 5 describes functional analysis of *cis*-regulatory elements that regulate spatiotemporal control of *snail* expression during *Drosophila* embryogenesis. Finally in Chapter 6, the significance of the work presented in this thesis is summarized and discussed.

References

Adams-Cioaba, M. A. and Min, J. (2009). Structure and function of histone methylation binding proteins. *Biochemistry and cell biology = Biochimie et biologie cellulaire* 87, 93-105.

Anderson, K. V., Bokla, L. and Nusslein-Volhard, C. (1985). Establishment of dorsal-ventral polarity in the Drosophila embryo: the induction of polarity by the Toll gene product. *Cell* 42, 791-798.

Bannister, A. J. and Kouzarides, T. (2005). Reversing histone methylation. *Nature* 436, 1103-1106.

Bannister, A. J., Schneider, R. and Kouzarides, T. (2002). Histone methylation: dynamic or static? *Cell* 109, 801-806.

Bannister, A. J., Schneider, R., Myers, F. A., Thorne, A. W., Crane-Robinson, C. and Kouzarides, T. (2005). Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *The Journal of biological chemistry* 280, 17732-17736.

Barrallo-Gimeno, A. and Nieto, M. A. (2005). The Snail genes as inducers of cell movement and survival: implications in development and cancer. *Development (Cambridge, England)* 132, 3151-3161.

Barrès, R., Osler, M. E., Yan, J., Rune, A., Fritz, T., Caidahl, K., Krook, A. and Zierath, J. R. (2009). Non-CpG Methylation of the PGC-1α Promoter through DNMT3B Controls Mitochondrial Density. *Cell Metabolism* 10, 189-198.

Batlle, E., Sancho, E., Franci, C., Dominguez, D., Monfar, M., Baulida, J. and Garcia De Herreros, A. (2000). The transcription factor snail is a repressor of E-cadherin gene expression in epithelial tumour cells. *Nature cell biology* 2, 84-89.

Baylies, M. K. and Bate, M. (1996). twist: a myogenic switch in Drosophila. *Science (New York, N.Y.)* 272, 1481-1484.

Becker, P. B. and Workman, J. L. (2013). Nucleosome remodeling and epigenetics. *Cold Spring Harbor perspectives in biology* 5.

Belle, J. I. and Nijnik, A. (2014). H2A-DUBbing the mammalian epigenome: expanding frontiers for histone H2A deubiquitinating enzymes in cell biology and physiology. *The international journal of biochemistry & cell biology* 50, 161-174.

Belvin, M. P. and Anderson, K. V. (1996). A conserved signaling pathway: the Drosophila toll-dorsal pathway. *Annual review of cell and developmental biology* 12, 393-416.

Bier, E., Jan, L. Y. and Jan, Y. N. (1990). rhomboid, a gene required for dorsoventral axis establishment and peripheral nervous system development in Drosophila melanogaster. *Genes & development* 4, 190-203.

Birke, M., Schreiner, S., Garcia-Cuellar, M. P., Mahr, K., Titgemeyer, F. and Slany, R. K. (2002). The MT domain of the proto-oncoprotein MLL binds to CpG-containing DNA and discriminates against methylation. *Nucleic acids research* 30, 958-965.

Black, J. C., Van Rechem, C. and Whetstine, J. R. (2012). Histone lysine methylation dynamics: establishment, regulation, and biological impact. *Molecular cell* 48, 491-507.

Bodmer, R., Jan, L. Y. and Jan, Y. N. (1990). A new homeobox-containing gene, msh-2, is transiently expressed early during mesoderm formation of Drosophila. *Development (Cambridge, England)* 110, 661-669.

Bonasio, R., Tu, S. and Reinberg, D. (2010). Molecular signals of epigenetic states. *Science (New York, N.Y.)* 330, 612-616.

Boutet, A., Esteban, M. A., Maxwell, P. H. and Nieto, M. A. (2007). Reactivation of Snail genes in renal fibrosis and carcinomas: a process of reversed embryogenesis? *Cell cycle (Georgetown, Tex.)* 6, 638-642.

Brunmeir, R., Lagger, S. and Seiser, C. (2009). Histone deacetylase HDAC1/HDAC2-controlled embryonic development and cell differentiation. *The International journal of developmental biology* 53, 275-289.

Buecker, C. and Wysocka, J. (2012). Enhancers as information integration hubs in development: lessons from genomics. *Trends in genetics : TIG* 28, 276-284.

Cano, A., Perez-Moreno, M. A., Rodrigo, I., Locascio, A., Blanco, M. J., del Barrio, M. G., Portillo, F. and Nieto, M. A. (2000). The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression. *Nature cell biology* 2, 76-83.

Casal, J. and Leptin, M. (1996). Identification of novel genes in Drosophila reveals the complex regulation of early gene activity in the mesoderm. *Proceedings of the National Academy of Sciences of the United States of America* 93, 10327-10332.

Cedar, H. and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nature reviews. Genetics* 10, 295-304.

Celic, I., Masumoto, H., Griffith, W. P., Meluh, P., Cotter, R. J., Boeke, J. D. and Verreault, A. (2006). The sirtuins hst3 and Hst4p preserve genome integrity by controlling histone h3 lysine 56 deacetylation. *Current biology : CB* 16, 1280-1289.

Chandy, M., Gutierrez, J. L., Prochasson, P. and Workman, J. L. (2006). SWI/SNF displaces SAGA-acetylated nucleosomes. *Eukaryotic cell* 5, 1738-1747.

Chasan, R. and Anderson, K. V. (1989). The role of easter, an apparent serine protease, in organizing the dorsal-ventral pattern of the Drosophila embryo. *Cell* 56, 391-400.

Cho, Y. S., Stevens, L. M. and Stein, D. (2010). Pipe-dependent ventral processing of Easter by Snake is the defining step in Drosophila embryo DV axis formation. *Current biology : CB* 20, 1133-1137.

Costa, M., Wilson, E. T. and Wieschaus, E. (1994). A putative cell signal encoded by the folded gastrulation gene coordinates cell shape changes during Drosophila gastrulation. *Cell* 76, 1075-1089.

Cripps, R. M., Black, B. L., Zhao, B., Lien, C. L., Schulz, R. A. and Olson, E. N. (1998).

The myogenic regulatory gene Mef2 is a direct target for transcriptional activation by Twist during Drosophila myogenesis. *Genes & development* 12, 422-434.

Cubenas-Potts, C. and Matunis, M. J. (2013). SUMO: a multifaceted modifier of chromatin structure and function. *Developmental cell* 24, 1-12.

Das, C., Lucia, M. S., Hansen, K. C. and Tyler, J. K. (2009). CBP/p300-mediated acetylation of histone H3 on lysine 56. *Nature* 459, 113-117.

Dawes-Hoang, R. E., Parmar, K. M., Christiansen, A. E., Phelps, C. B., Brand, A. H. and Wieschaus, E. F. (2005). folded gastrulation, cell shape change and the control of myosin localization. *Development (Cambridge, England)* 132, 4165-4178.

de Ruijter, A. J., van Gennip, A. H., Caron, H. N., Kemp, S. and van Kuilenburg, A. B. (2003). Histone deacetylases (HDACs): characterization of the classical HDAC family. *The Biochemical journal* 370, 737-749.

Del Rizzo, P. A. and Trievel, R. C. (2011). Substrate and product specificities of SET domain methyltransferases. *Epigenetics : official journal of the DNA Methylation Society* 6, 1059-1067.

---- (2014). Molecular basis for substrate recognition by lysine methyltransferases and demethylases. *Biochimica et biophysica acta*.

DeLotto, Y. and DeLotto, R. (1998). Proteolytic processing of the Drosophila Spatzle protein by easter generates a dimeric NGF-like molecule with ventralising activity. *Mechanisms of development* 72, 141-148.

Downs, J. A., Allard, S., Jobin-Robitaille, O., Javaheri, A., Auger, A., Bouchard, N., Kron, S. J., Jackson, S. P. and Cote, J. (2004). Binding of chromatin-modifying activities to phosphorylated histone H2A at DNA damage sites. *Molecular cell* 16, 979-990.

Drier, E. A., Huang, L. H. and Steward, R. (1999). Nuclear import of the Drosophila Rel protein Dorsal is regulated by phosphorylation. *Genes & development* 13, 556-568.

Driscoll, R., Hudson, A. and Jackson, S. P. (2007). Yeast Rtt109 promotes genome stability by acetylating histone H3 on lysine 56. *Science (New York, N.Y.)* 315, 649-652.

Edwards, D. N., Towb, P. and Wasserman, S. A. (1997). An activity-dependent network of interactions links the Rel protein Dorsal with its cytoplasmic regulators. *Development (Cambridge, England)* 124, 3855-3864.

Eissenberg, J. C. (2012). Structural biology of the chromodomain: Form and function. *Gene* 496, 69-78.

Euskirchen, G., Auerbach, R. K. and Snyder, M. (2012). SWI/SNF chromatin-remodeling factors: multiscale analyses and diverse functions. *The Journal of biological chemistry* 287, 30897-30905.

Fang, J., Chen, T., Chadwick, B., Li, E. and Zhang, Y. (2004). Ring1b-mediated H2A ubiquitination associates with inactive X chromosomes and is involved in initi-

ation of X inactivation. *The Journal of biological chemistry* 279, 52812-52815.

Ferguson-Smith, A. C. (2011). Genomic imprinting: the emergence of an epigenetic paradigm. *Nature reviews. Genetics* 12, 565-575.

Ferguson-Smith, A. C. and Greally, J. M. (2007). Epigenetics: Perceptive enzymes. *Nature* 449, 148-149.

Ferguson, E. L. and Anderson, K. V. (1992a). Decapentaplegic acts as a morphogen to organize dorsal-ventral pattern in the Drosophila embryo. *Cell* 71, 451-461.

---- (1992b). Localized enhancement and repression of the activity of the TGF-beta family member, decapentaplegic, is necessary for dorsal-ventral pattern formation in the Drosophila embryo. *Development (Cambridge, England)* 114, 583-597.

Francois, V., Solloway, M., O'Neill, J. W., Emery, J. and Bier, E. (1994). Dorsal-ventral patterning of the Drosophila embryo depends on a putative negative growth factor encoded by the short gastrulation gene. *Genes & development* 8, 2602-2616.

Frohnhofer, H. G., Lehmann, R. and Nusslein-Volhard, C. (1986). Manipulating the anteroposterior pattern of the Drosophila embryo. *Journal of embryology and experimental morphology* 97 Suppl, 169-179.

Fuchs, G. and Oren, M. (2014). Writing and reading H2B monoubiquitylation. *Biochimica et biophysica acta* 1839, 694-701.

Fuda, N. J., Ardehali, M. B. and Lis, J. T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* 461, 186-192.

Garcia, M. and Stathopoulos, A. (2011). Lateral gene expression in Drosophila early embryos is supported by Grainyhead-mediated activation and tiers of dorsally-localized repression. *PloS one* 6, e29172.

Gayatri, S. and Bedford, M. T. (2014). Readers of histone methylarginine marks. *Biochimica et biophysica acta* 1839, 702-710.

Geisler, R., Bergmann, A., Hiromi, Y. and Nusslein-Volhard, C. (1992). cactus, a gene involved in dorsoventral pattern formation of Drosophila, is related to the I kappa B gene family of vertebrates. *Cell* 71, 613-621.

Gerhold, C. B. and Gasser, S. M. (2014). INO80 and SWR complexes: relating structure to function in chromatin remodeling. *Trends in cell biology*.

Ghirlando, R. and Felsenfeld, G. (2013). Chromatin structure outside and inside the nucleus. *Biopolymers* 99, 225-232.

Giannattasio, M., Lazzaro, F., Plevani, P. and Muzi-Falconi, M. (2005). The DNA damage checkpoint response requires histone H2B ubiquitination by Rad6-Bre1 and H3 methylation by Dot1. *The Journal of biological chemistry* 280, 9879-9886.

Glatt, S., Alfieri, C. and Muller, C. W. (2011). Recognizing and remodeling the nucleosome. *Current opinion in structural biology* 21, 335-341.

Gopalakrishnan, S., Sullivan, B. A., Trazzi, S., Della Valle, G. and Robertson, K. D. (2009). DNMT3B interacts with constitutive centromere protein CENP-C to

modulate DNA methylation and the histone code at centromeric regions. *Human Molecular Genetics* 18, 3178-3193.

Gregoretti, I., Lee, Y.-M. and Goodson, H. V. (2004). Molecular Evolution of the Histone Deacetylase Family: Functional Implications of Phylogenetic Analysis. *Journal of Molecular Biology* 338, 17-31.

Grigoryev, S. A. and Woodcock, C. L. (2012). Chromatin organization - the 30 nm fiber. *Experimental cell research* 318, 1448-1455.

Grunstein, M. (1997). Histone acetylation in chromatin structure and transcription. *Nature* 389, 349-352.

Guo, J. U., Su, Y., Shin, J. H., Shin, J., Li, H., Xie, B., Zhong, C., Hu, S., Le, T., Fan, G., et al. (2014). Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat Neurosci* 17, 215-222.

Gurard-Levin, Z. A. and Almouzni, G. (2014). Histone modifications and a choice of variant: a language that helps the genome express itself. *F1000prime reports* 6, 76.

Halfon, M. S., Hashimoto, C. and Keshishian, H. (1995). The Drosophila toll gene functions zygotically and is necessary for proper motoneuron and muscle development. *Developmental biology* 169, 151-167.

Han, J., Zhou, H., Li, Z., Xu, R. M. and Zhang, Z. (2007). Acetylation of lysine 56 of histone H3 catalyzed by RTT109 and regulated by ASF1 is required for replisome integrity. *The Journal of biological chemistry* 282, 28587-28596.

Harfe, B. D., Vaz Gomes, A., Kenyon, C., Liu, J., Krause, M. and Fire, A. (1998). Analysis of a Caenorhabditis elegans Twist homolog identifies conserved and divergent aspects of mesodermal patterning. *Genes & development* 12, 2623-2635.

Hargreaves, D. C. and Crabtree, G. R. (2011). ATP-dependent chromatin remodeling: genetics, genomics and mechanisms. *Cell research* 21, 396-420.

Hecht, P. M. and Anderson, K. V. (1993). Genetic characterization of tube and pelle, genes required for signaling between Toll and dorsal in the specification of the dorsal-ventral pattern of the Drosophila embryo. *Genetics* 135, 405-417.

Hellman, A. and Chess, A. (2007). Gene Body-Specific Methylation on the Active X Chromosome. *Science (New York, N.Y.)* 315, 1141-1143.

Hemavathy, K., Guru, S. C., Harris, J., Chen, J. D. and Ip, Y. T. (2000). Human Slug is a repressor that localizes to sites of active transcription. *Molecular and cellular biology* 20, 5087-5095.

Hemavathy, K., Meng, X. and Ip, Y. T. (1997). Differential regulation of gastrulation and neuroectodermal gene expression by Snail in the Drosophila embryo. *Development (Cambridge, England)* 124, 3683-3691.

Hong, C. C. and Hashimoto, C. (1995). An unusual mosaic protein with a protease domain, encoded by the nudel gene, is involved in defining embryonic dorsoventral polarity in Drosophila. *Cell* 82, 785-794.

Hsu, J. Y., Sun, Z. W., Li, X., Reuben, M., Tatchell, K., Bishop, D. K., Grushcow, J. M., Brame, C. J., Caldwell, J. A., Hunt, D. F., et al. (2000). Mitotic phosphorylation of histone H3 is governed by Ipl1/aurora kinase and Glc7/PP1 phosphatase in budding yeast and nematodes. *Cell* 102, 279-291.

Huang, Y., Fang, J., Bedford, M. T., Zhang, Y. and Xu, R. M. (2006). Recognition of histone H3 lysine-4 methylation by the double tudor domain of JMJD2A. *Science (New York, N.Y.)* 312, 748-751.

Huen, M. S., Grant, R., Manke, I., Minn, K., Yu, X., Yaffe, M. B. and Chen, J. (2007). RNF8 transduces the DNA-damage signal via histone ubiquitylation and checkpoint protein assembly. *Cell* 131, 901-914.

Hyland, E. M., Cosgrove, M. S., Molina, H., Wang, D., Pandey, A., Cottee, R. J. and Boeke, J. D. (2005). Insights into the role of histone H3 and histone H4 core modifiable residues in Saccharomyces cerevisiae. *Molecular and cellular biology* 25, 10060-10070.

Ip, Y. T., Kraut, R., Levine, M. and Rushlow, C. A. (1991). The dorsal morphogen is a sequence-specific DNA-binding protein that interacts with a long-range repression element in Drosophila. *Cell* 64, 439-446.

Ip, Y. T., Park, R. E., Kosman, D., Bier, E. and Levine, M. (1992a). The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive neuroectoderm of the Drosophila embryo. *Genes & development* 6, 1728-1739.

Ip, Y. T., Park, R. E., Kosman, D., Yazdanbakhsh, K. and Levine, M. (1992b). dorsal-twist interactions establish snail expression in the presumptive mesoderm of the Drosophila embryo. *Genes & development* 6, 1518-1530.

Jacobs, S. A. and Khorasanizadeh, S. (2002). Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science (New York, N.Y.)* 295, 2080-2083.

Jang, I. H., Chosa, N., Kim, S. H., Nam, H. J., Lemaitre, B., Ochiai, M., Kambris, Z., Brun, S., Hashimoto, C., Ashida, M., et al. (2006). A Spatzle-processing enzyme required for toll signaling activation in Drosophila innate immunity. *Developmental cell* 10, 45-55.

Jason, L. J., Moore, S. C., Lewis, J. D., Lindsey, G. and Ausio, J. (2002). Histone ubiquitination: a tagging tail unfolds? *BioEssays : news and reviews in molecular, cellular and developmental biology* 24, 166-174.

Jazwinska, A., Rushlow, C. and Roth, S. (1999). The role of brinker in mediating the graded response to Dpp in early Drosophila embryos. *Development (Cambridge, England)* 126, 3323-3334.

Jiang, J., Kosman, D., Ip, Y. T. and Levine, M. (1991). The dorsal morphogen gradient regulates the mesoderm determinant twist in early Drosophila embryos. *Genes & development* 5, 1881-1891.

Jiang, J. and Levine, M. (1993). Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen. *Cell* 72, 741-752.

Jimenez, F., Martin-Morris, L. E., Velasco, L., Chu, H., Sierra, J., Rosen, D. R. and White, K. (1995). vnd, a gene required for early neurogenesis of Drosophila, encodes a homeodomain protein. *The EMBO journal* 14, 3487-3495.

Jin, B., Tao, Q., Peng, J., Soo, H. M., Wu, W., Ying, J., Fields, C. R., Delmas, A. L., Liu, X., Qiu, J., et al. (2008). DNA methyltransferase 3B (DNMT3B) mutations in ICF syndrome lead to altered epigenetic modifications and aberrant expression of genes regulating development, neurogenesis and immune function. *Human Molecular Genetics* 17, 690-709.

Johnson, D. G. and Dent, S. Y. (2013). Chromatin: receiver and quarterback for cellular signals. *Cell* 152, 685-689.

Joo, H. Y., Zhai, L., Yang, C., Nie, S., Erdjument-Bromage, H., Tempst, P., Chang, C. and Wang, H. (2007). Regulation of cell cycle progression and gene expression by H2A deubiquitination. *Nature* 449, 1068-1072.

Kalb, R., Latwiel, S., Baymaz, H. I., Jansen, P. W., Muller, C. W., Vermeulen, M. and Muller, J. (2014). Histone H2A monoubiquitination promotes histone H3 methylation in Polycomb repression. *Nature structural & molecular biology* 21, 569-571.

Kar, S., Deb, M., Sengupta, D., Shilpi, A., Parbin, S., Torrisani, J., Pradhan, S. and Patra, S. K. (2012). An insight into the various regulatory mechanisms modulating human DNA methyltransferase 1 stability and function. *Epigenetics : official journal of the DNA Methylation Society* 7, 994-1007.

Kehle, J., Beuchle, D., Treuheit, S., Christen, B., Kennison, J. A., Bienz, M. and Muller, J. (1998). dMi-2, a hunchback-interacting protein that functions in polycomb repression. *Science (New York, N.Y.)* 282, 1897-1900.

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., Ward, L. D., Birney, E., Crawford, G. E., Dekker, J., et al. (2014). Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* 111, 6131-6138.

Kidd, S. (1992). Characterization of the Drosophila cactus locus and analysis of interactions between cactus and dorsal proteins. *Cell* 71, 623-635.

Klambt, C., Knust, E., Tietze, K. and Campos-Ortega, J. A. (1989). Closely related transcripts encoded by the neurogenic gene complex enhancer of split of Drosophila melanogaster. *The EMBO journal* 8, 203-210.

Klingler, M., Erdelyi, M., Szabad, J. and Nusslein-Volhard, C. (1988). Function of torso in determining the terminal anlagen of the Drosophila embryo. *Nature* 335, 275-277.

Klose, R. J., Yamane, K., Bae, Y., Zhang, D., Erdjument-Bromage, H., Tempst, P., Wong, J. and Zhang, Y. (2006). The transcriptional repressor JHDM3A demethylates trimethyl histone H3 lysine 9 and lysine 36. *Nature* 442, 312-316.

Kolsch, V., Seher, T., Fernandez-Ballester, G. J., Serrano, L. and Leptin, M. (2007). Control of Drosophila gastrulation by apical localization of adherens junctions and RhoGEF2. *Science (New York, N.Y.)* 315, 384-386.

Kosman, D., Ip, Y. T., Levine, M. and Arora, K. (1991). Establishment of the mesoderm-neuroectoderm boundary in the Drosophila embryo. *Science (New York, N.Y.)* 254, 118-122.

Kouzarides, T. (2007). Chromatin Modifications and Their Function. *Cell* 128, 693-705.

Krogan, N. J., Keogh, M. C., Datta, N., Sawa, C., Ryan, O. W., Ding, H., Haw, R. A., Pootoolal, J., Tong, A., Canadien, V., et al. (2003a). A Snf2 family ATPase complex required for recruitment of the histone H2A variant Htz1. *Molecular cell* 12, 1565-1576.

Krogan, N. J., Kim, M., Tong, A., Golshani, A., Cagney, G., Canadien, V., Richards, D. P., Beattie, B. K., Emili, A., Boone, C., et al. (2003b). Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. *Molecular and cellular biology* 23, 4207-4218.

Lalonde, M. E., Cheng, X. and Cote, J. (2014). Histone target selection within chromatin: an exemplary case of teamwork. *Genes & development* 28, 1029-1041.

Law, J. A. and Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews. Genetics* 11, 204-220.

Lee, J. S., Shukla, A., Schneider, J., Swanson, S. K., Washburn, M. P., Florens, L., Bhaumik, S. R. and Shilatifard, A. (2007). Histone crosstalk between H2B monoubiquitination and H3 methylation mediated by COMPASS. *Cell* 131, 1084-1096.

Lee, K. K. and Workman, J. L. (2007). Histone acetyltransferase complexes: one size doesn't fit all. *Nature reviews. Molecular cell biology* 8, 284-295.

Lee, M. G., Wynder, C., Bochar, D. A., Hakimi, M. A., Cooch, N. and Shiekhattar, R. (2006). Functional interplay between histone demethylase and deacetylase enzymes. *Molecular and cellular biology* 26, 6395-6402.

Lehmann, R. and Nusslein-Volhard, C. (1986). Abdominal segmentation, pole cell formation, and embryonic polarity require the localized activity of oskar, a maternal gene in Drosophila. *Cell* 47, 141-152.

Lelli, K. M., Slattery, M. and Mann, R. S. (2012). Disentangling the many layers of eukaryotic transcriptional regulation. *Annual review of genetics* 46, 43-68.

Lemaitre, B., Nicolas, E., Michaut, L., Reichhart, J. M. and Hoffmann, J. A. (1996). The dorsoventral regulatory gene cassette spatzle/Toll/cactus controls the potent antifungal response in Drosophila adults. *Cell* 86, 973-983.

Leptin, M. (1991). twist and snail as positive and negative regulators during Drosophila mesoderm development. *Genes & development* 5, 1568-1576.

Leptin, M. and Grunewald, B. (1990). Cell shape changes during gastrulation in Drosophila. *Development (Cambridge, England)* 110, 73-84.

Levine, M. and Tjian, R. (2003). Transcription regulation and animal diversity. *Nature* 424, 147-151.

Levo, M. and Segal, E. (2014). In pursuit of design principles of regulatory sequences. *Nature reviews. Genetics* 15, 453-468.

Liberman, L. M. and Stathopoulos, A. (2009). Design flexibility in cis-regulatory control of gene expression: synthetic and comparative evidence. *Developmental biology* 327, 578-589.

Lilly, B., Galewsky, S., Firulli, A. B., Schulz, R. A. and Olson, E. N. (1994). D-MEF2: A MADS box transcription factor expressed in differentiating mesoderm and muscle cell lineages during Drosophila embryogenesis. *Proceedings of the National Academy of Sciences of the United States of America* 91, 5662-5666.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462, 315-322.

Luger, K. (2003). Structure and dynamic behavior of nucleosomes. *Current opinion in genetics & development* 13, 127-135.

Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. and Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* 389, 251-260.

Maas, N. L., Miller, K. M., DeFazio, L. G. and Toczyski, D. P. (2006). Cell cycle and checkpoint regulation of histone H3 K56 acetylation by Hst3 and Hst4. *Molecular cell* 23, 109-119.

Macfarlan, T. S., Gifford, W. D., Driscoll, S., Lettieri, K., Rowe, H. M., Bonanomi, D., Firth, A., Singer, O., Trono, D. and Pfaff, S. L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* 487, 57-63.

Markstein, M., Markstein, P., Markstein, V. and Levine, M. S. (2002). Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proceedings of the National Academy of Sciences of the United States of America* 99, 763-768.

Markstein, M., Zinzen, R., Markstein, P., Yee, K. P., Erives, A., Stathopoulos, A. and Levine, M. (2004). A regulatory code for neurogenic gene expression in the Drosophila embryo. *Development (Cambridge, England)* 131, 2387-2394.

Massari, M. E. and Murre, C. (2000). Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Molecular and cellular biology* 20, 429-440.

Masumoto, H., Hawke, D., Kobayashi, R. and Verreault, A. (2005). A role for cell-cycle-regulated histone H3 lysine 56 acetylation in the DNA damage response. *Nature* 436, 294-298.

Migliori, V., Mapelli, M. and Guccione, E. (2012). On WD40 proteins: propelling our knowledge of transcriptional control? *Epigenetics : official journal of the DNA Methylation Society* 7, 815-822.

Mizuguchi, G., Shen, X., Landry, J., Wu, W. H., Sen, S. and Wu, C. (2004). ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling

complex. *Science (New York, N.Y.)* 303, 343-348.

Morel, V. and Schweisguth, F. (2000). Repression by suppressor of hairless and activation by Notch are required to define a single row of single-minded expressing cells in the Drosophila embryo. *Genes & development* 14, 377-388.

Morgan, M. M. and Mahowald, A. P. (1996). Multiple signaling pathways establish both the individuation and the polarity of the oocyte follicle in Drosophila. *Archives of insect biochemistry and physiology* 33, 211-230.

Morris, S. A., Shibata, Y., Noma, K., Tsukamoto, Y., Warren, E., Temple, B., Grewal, S. I. and Strahl, B. D. (2005). Histone H3 K36 methylation is associated with transcription elongation in Schizosaccharomyces pombe. *Eukaryotic cell* 4, 1446-1454.

Morrison, A. J., Highland, J., Krogan, N. J., Arbel-Eden, A., Greenblatt, J. F., Haber, J. E. and Shen, X. (2004). INO80 and gamma-H2AX interaction links ATP-dependent chromatin remodeling to DNA damage repair. *Cell* 119, 767-775.

Mosammaparast, N. and Shi, Y. (2010). Reversal of histone methylation: biochemical and molecular mechanisms of histone demethylases. *Annual review of biochemistry* 79, 155-179.

Moser, M. A., Hagelkruys, A. and Seiser, C. (2014). Transcription and beyond: the role of mammalian class I lysine deacetylases. *Chromosoma* 123, 67-78.

Moussian, B. and Roth, S. (2005). Dorsoventral axis formation in the Drosophila embryo--shaping and transducing a morphogen gradient. *Current biology : CB* 15, R887-899.

Musselman, C. A. and Kutateladze, T. G. (2011). Handpicking epigenetic marks with PHD fingers. *Nucleic acids research* 39, 9061-9071.

Nagy, Z. and Tora, L. (2007). Distinct GCN5/PCAF-containing complexes function as co-activators and are involved in transcription factor and global histone acetylation. *Oncogene* 26, 5341-5357.

Nambu, J. R., Franks, R. G., Hu, S. and Crews, S. T. (1990). The single-minded gene of Drosophila is required for the expression of genes important for the development of CNS midline cells. *Cell* 63, 63-75.

Nan, X., Ng, H. H., Johnson, C. A., Laherty, C. D., Turner, B. M., Eisenman, R. N. and Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* 393, 386-389.

Narlikar, G. J., Sundaramoorthy, R. and Owen-Hughes, T. (2013). Mechanisms and functions of ATP-dependent chromatin-remodeling enzymes. *Cell* 154, 490-503.

Nathan, D., Ingvarsdottir, K., Sterner, D. E., Bylebyl, G. R., Dokmanovic, M., Dorsey, J. A., Whelan, K. A., Krsmanovic, M., Lane, W. S., Meluh, P. B., et al. (2006). Histone sumoylation is a negative regulator in Saccharomyces cerevisiae and shows dynamic interplay with positive-acting histone modifications. *Genes & development* 20, 966-976.

Nathan, D., Sterner, D. E. and Berger, S. L. (2003). Histone modifications: Now summoning sumoylation. *Proceedings of the National Academy of Sciences of the United States of America* 100, 13118-13120.

Ng, H. H., Ciccone, D. N., Morshead, K. B., Oettinger, M. A. and Struhl, K. (2003). Lysine-79 of histone H3 is hypomethylated at silenced loci in yeast and mammalian cells: a potential mechanism for position-effect variegation. *Proceedings of the National Academy of Sciences of the United States of America* 100, 1820-1825.

Ng, H. H., Xu, R. M., Zhang, Y. and Struhl, K. (2002). Ubiquitination of histone H2B by Rad6 is required for efficient Dot1-mediated methylation of histone H3 lysine 79. *The Journal of biological chemistry* 277, 34655-34657.

Nielsen, P. R., Nietlispach, D., Mott, H. R., Callaghan, J., Bannister, A., Kouzarides, T., Murzin, A. G., Murzina, N. V. and Laue, E. D. (2002). Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. *Nature* 416, 103-107.

Nieto, M. A. (2002). The snail superfamily of zinc-finger transcription factors. *Nature reviews. Molecular cell biology* 3, 155-166.

Nijman, S. M., Luna-Vargas, M. P., Velds, A., Brummelkamp, T. R., Dirac, A. M., Sixma, T. K. and Bernards, R. (2005). A genomic and functional inventory of deubiquitinating enzymes. *Cell* 123, 773-786.

Nusslein-Volhard, C., Frohnhofer, H. G. and Lehmann, R. (1987). Determination of anteroposterior polarity in Drosophila. *Science (New York, N.Y.)* 238, 1675-1681.

Nusslein-Volhard, C. and Wieschaus, E. (1980). Mutations affecting segment number and polarity in Drosophila. *Nature* 287, 795-801.

Ong, C. T. and Corces, V. G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature reviews. Genetics* 12, 283-293.

Ozdemir, A., Spicuglia, S., Lasonder, E., Vermeulen, M., Campsteijn, C., Stunnenberg, H. G. and Logie, C. (2005). Characterization of lysine 56 of histone H3 as an acetylation site in Saccharomyces cerevisiae. *The Journal of biological chemistry* 280, 25949-25952.

Pan, D. J., Huang, J. D. and Courey, A. J. (1991). Functional analysis of the Drosophila twist promoter reveals a dorsal-binding ventral activator region. *Genes & development* 5, 1892-1901.

Patel, D. J. and Wang, Z. (2013). Readout of epigenetic modifications. *Annual review of biochemistry* 82, 81-118.

Peinado, H., Olmeda, D. and Cano, A. (2007). Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype? *Nature reviews. Cancer* 7, 415-428.

Perez-Losada, J., Sanchez-Martin, M., Perez-Caro, M., Perez-Mancera, P. A. and Sanchez-Garcia, I. (2003). The radioresistance biological function of the SCF//kit signaling pathway is mediated by the zinc-finger transcription factor Slug. *Oncogene* 22, 4205-4211.

Peri, F., Technau, M. and Roth, S. (2002). Mechanisms of Gurken-dependent pipe regulation and the robustness of dorsoventral patterning in Drosophila. *Development (Cambridge, England)* 129, 2965-2975.

Petty, E. and Pillus, L. (2013). Balancing chromatin remodeling and histone modifications in transcription. *Trends in genetics : TIG* 29, 621-629.

Plank, J. L. and Dean, A. (2014). Enhancer function: mechanistic and genome-wide insights come together. *Molecular cell* 55, 5-14.

Pray-Grant, M. G., Daniel, J. A., Schieltz, D., Yates, J. R., 3rd and Grant, P. A. (2005). Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. *Nature* 433, 434-438.

Prendergast, G. and Ziff, E. (1991). Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. *Science (New York, N.Y.)* 251, 186-189.

Qian, C. and Zhou, M. M. (2006). SET domain protein lysine methyltransferases: Structure, specificity and catalysis. *Cellular and molecular life sciences : CMLS* 63, 2755-2763.

Qiu, P., Pan, P. C. and Govind, S. (1998). A role for the Drosophila Toll/Cactus pathway in larval hematopoiesis. *Development (Cambridge, England)* 125, 1909-1920.

Raisner, R. M., Hartley, P. D., Meneghini, M. D., Bao, M. Z., Liu, C. L., Schreiber, S. L., Rando, O. J. and Madhani, H. D. (2005). Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin. *Cell* 123, 233-248.

Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P. and Jaenisch, R. (2000). Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences* 97, 5237-5242.

Reach, M., Galindo, R. L., Towb, P., Allen, J. L., Karin, M. and Wasserman, S. A. (1996). A gradient of cactus protein degradation establishes dorsoventral polarity in the Drosophila embryo. *Developmental biology* 180, 353-364.

Recht, J., Tsubota, T., Tanny, J. C., Diaz, R. L., Berger, J. M., Zhang, X., Garcia, B. A., Shabanowitz, J., Burlingame, A. L., Hunt, D. F., et al. (2006). Histone chaperone Asf1 is required for histone H3 lysine 56 acetylation, a modification associated with S phase in mitosis and meiosis. *Proceedings of the National Academy of Sciences of the United States of America* 103, 6988-6993.

Reeves, G. T. and Stathopoulos, A. (2009). Graded dorsal and differential gene regulation in the Drosophila embryo. *Cold Spring Harbor perspectives in biology* 1, a000836.

Reuter, R. and Leptin, M. (1994). Interacting functions of snail, twist and huckebein during the early development of germ layers in Drosophila. *Development (Cambridge, England)* 120, 1137-1150.

Robertson, K. D., Ait-Si-Ali, S., Yokochi, T., Wade, P. A., Jones, P. L. and Wolffe, A. P. (2000). DNMT1 forms a complex with Rb, E2F1 and HDAC1 and represses transcription from E2F-responsive promoters. *Nature genetics* 25, 338-342.

Robzyk, K., Recht, J. and Osley, M. A. (2000). Rad6-dependent ubiquitination of histone H2B in yeast. *Science (New York, N.Y.)* 287, 501-504.

Rogakou, E. P., Pilch, D. R., Orr, A. H., Ivanova, V. S. and Bonner, W. M. (1998). DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *The Journal of biological chemistry* 273, 5858-5868.

Rose, N. R. and Klose, R. J. (2014). Understanding the relationship between DNA methylation and histone lysine methylation. *Biochimica et biophysica acta*.

Rossetto, D., Avvakumov, N. and Cote, J. (2012). Histone phosphorylation: a chromatin modification involved in diverse nuclear events. *Epigenetics : official journal of the DNA Methylation Society* 7, 1098-1108.

Roth, S., Hiromi, Y., Godt, D. and Nusslein-Volhard, C. (1991). cactus, a maternal gene required for proper formation of the dorsoventral morphogen gradient in Drosophila embryos. *Development (Cambridge, England)* 112, 371-388.

Roth, S., Stein, D. and Nusslein-Volhard, C. (1989). A gradient of nuclear localization of the dorsal protein determines dorsoventral pattern in the Drosophila embryo. *Cell* 59, 1189-1202.

Rothbart, S. B. and Strahl, B. D. (2014). Interpreting the language of histone and DNA modifications. *Biochimica et biophysica acta* 1839, 627-643.

Rountree, M. R., Bachman, K. E. and Baylin, S. B. (2000). DNMT1 binds HDAC2 and a new co-repressor, DMAP1, to form a complex at replication foci. *Nature genetics* 25, 269-277.

Rudolph, T., Beuch, S. and Reuter, G. (2013). Lysine-specific histone demethylase LSD1 and the dynamic control of chromatin. *Biological chemistry* 394, 1019-1028.

Rushlow, C. and Levine, M. (1990). Role of the zerknullt gene in dorsal-ventral pattern formation in Drosophila. *Advances in genetics* 27, 277-307.

Rushlow, C. A., Han, K., Manley, J. L. and Levine, M. (1989). The graded distribution of the dorsal morphogen is initiated by selective nuclear transport in Drosophila. *Cell* 59, 1165-1177.

Rushlow, C. A. and Shvartsman, S. Y. (2012). Temporal dynamics, spatial range, and transcriptional interpretation of the Dorsal morphogen gradient. *Current opinion in genetics & development* 22, 542-546.

Saeed, S., Quintin, J., Kerstens, H. H. D., Rao, N. A., Aghajanirefah, A., Matarese, F., Cheng, S.-C., Ratter, J., Berentsen, K., van der Ent, M. A., et al. (2014). Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science (New York, N.Y.)* 345.

Sanchez, R., Meslamani, J. and Zhou, M. M. (2014). The bromodomain: from epigenome reader to druggable target. *Biochimica et biophysica acta* 1839, 676-685.

Sandmann, T., Girardot, C., Brehme, M., Tongprasit, W., Stolc, V. and Furlong, E. E. (2007). A core transcriptional network for early mesoderm development in Drosophila melanogaster. *Genes & development* 21, 436-449.

Sandmann, T., Jensen, L. J., Jakobsen, J. S., Karzynski, M. M., Eichenlaub, M. P., Bork, P. and Furlong, E. E. (2006). A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Developmental cell* 10, 797-807.

Schaefer, M. and Lyko, F. (2010). Solving the Dnmt2 enigma. *Chromosoma* 119, 35-40.

Schneider, D. S., Jin, Y., Morisato, D. and Anderson, K. V. (1994). A processed form of the Spatzle protein defines dorsal-ventral polarity in the Drosophila embryo. *Development (Cambridge, England)* 120, 1243-1250.

Schneider, J., Bajwa, P., Johnson, F. C., Bhaumik, S. R. and Shilatifard, A. (2006). Rtt109 is required for proper H3K56 acetylation: a chromatin mark associated with the elongating RNA polymerase II. *The Journal of biological chemistry* 281, 37270-37274.

Schotta, G., Ebert, A., Krauss, V., Fischer, A., Hoffmann, J., Rea, S., Jenuwein, T., Dorn, R. and Reuter, G. (2002). Central role of Drosophila SU(VAR)3-9 in histone H3-K9 methylation and heterochromatic gene silencing. *The EMBO journal* 21, 1121-1131.

Schubeler, D., MacAlpine, D. M., Scalzo, D., Wirbelauer, C., Kooperberg, C., van Leeuwen, F., Gottschling, D. E., O'Neill, L. P., Turner, B. M., Delrow, J., et al. (2004). The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes & development* 18, 1263-1271.

Schupbach, T. and Wieschaus, E. (1986). Germline autonomy of maternal-effect mutations altering the embryonic body pattern of Drosophila. *Developmental biology* 113, 443-448.

Sen, J., Goltz, J. S., Stevens, L. and Stein, D. (1998). Spatially restricted expression of pipe in the Drosophila egg chamber defines embryonic dorsal-ventral polarity. *Cell* 95, 471-481.

Shahbazian, M. D. and Grunstein, M. (2007). Functions of site-specific histone acetylation and deacetylation. *Annual review of biochemistry* 76, 75-100.

Shi, Y., Lan, F., Matson, C., Mulligan, P., Whetstine, J. R., Cole, P. A., Casero, R. A. and Shi, Y. (2004). Histone Demethylation Mediated by the Nuclear Amine Oxidase Homolog LSD1. *Cell* 119, 941-953.

Shiio, Y. and Eisenman, R. N. (2003). Histone sumoylation is associated with transcriptional repression. *Proceedings of the National Academy of Sciences of the United States of America* 100, 13225-13230.

Shishido, E., Higashijima, S., Emori, Y. and Saigo, K. (1993). Two FGF-receptor homologues of Drosophila: one is expressed in mesodermal primordium in early embryos. *Development (Cambridge, England)* 117, 751-761.

Shlyueva, D., Stampfel, G. and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature reviews. Genetics* 15, 272-286.

Shmakova, A., Batie, M., Druker, J. and Rocha, S. (2014). Chromatin and oxygen sensing in the context of JmjC histone demethylases. *The Biochemical journal* 462,

385-395.

Shroff, R., Arbel-Eden, A., Pilch, D., Ira, G., Bonner, W. M., Petrini, J. H., Haber, J. E. and Lichten, M. (2004). Distribution and dynamics of chromatin modification induced by a defined DNA double-strand break. *Current biology : CB* 14, 1703-1711.

Shukla, A. and Bhaumik, S. R. (2007). H2B-K123 ubiquitination stimulates RNAPII elongation independent of H3-K4 methylation. *Biochemical and biophysical research communications* 359, 214-220.

Simpson, P. (1983). Maternal-Zygotic Gene Interactions during Formation of the Dorsoventral Pattern in Drosophila Embryos. *Genetics* 105, 615-632.

Spitz, F. and Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nature reviews. Genetics* 13, 613-626.

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E. J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 480, 490-495.

Stanisavljevic, J., Porta-de-la-Riva, M., Batlle, R., de Herreros, A. G. and Baulida, J. (2011). The p65 subunit of NF-κB and PARP1 assist Snail1 in activating fibronectin transcription. *Journal of Cell Science* 124, 4161-4171.

Stathopoulos, A. and Levine, M. (2005). Genomic regulatory networks and animal development. *Developmental cell* 9, 449-462.

Stathopoulos, A., Tam, B., Ronshaugen, M., Frasch, M. and Levine, M. (2004). pyramus and thisbe: FGF genes that pattern the mesoderm of Drosophila embryos. *Genes & development* 18, 687-699.

Stathopoulos, A., Van Drenth, M., Erives, A., Markstein, M. and Levine, M. (2002). Whole-genome analysis of dorsal-ventral patterning in the Drosophila embryo. *Cell* 111, 687-701.

Stein, D. and Nusslein-Volhard, C. (1992). Multiple extracellular activities in Drosophila egg perivitelline fluid are required for establishment of embryonic dorsal-ventral polarity. *Cell* 68, 429-440.

Stein, D., Roth, S., Vogelsang, E. and Nusslein-Volhard, C. (1991). The polarity of the dorsoventral axis in the Drosophila embryo is defined by an extracellular signal. *Cell* 65, 725-735.

Stiff, T., O'Driscoll, M., Rief, N., Iwabuchi, K., Lobrich, M. and Jeggo, P. A. (2004). ATM and DNA-PK function redundantly to phosphorylate H2AX after exposure to ionizing radiation. *Cancer research* 64, 2390-2396.

Suganuma, T. and Workman, J. L. (2011). Signals and combinatorial functions of histone modifications. *Annual review of biochemistry* 80, 473-499.

Sun, Z. W. and Allis, C. D. (2002). Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast. *Nature* 418, 104-108.

Swygert, S. G. and Peterson, C. L. (2014). Chromatin dynamics: interplay between remodeling enzymes and histone modifications. *Biochimica et biophysica acta* 1839, 728-736.

Tan, Y., Xue, Y., Song, C. and Grunstein, M. (2013). Acetylated histone H3K56 interacts with Oct4 to promote mouse embryonic stem cell pluripotency. *Proceedings of the National Academy of Sciences of the United States of America* 110, 11493-11498.

Tessarz, P. and Kouzarides, T. (2014). Histone core modifications regulating nucleosome structure and dynamics. *Nat Rev Mol Cell Biol* 15, 703-708.

Thiery, J. P. and Sleeman, J. P. (2006). Complex networks orchestrate epithelial-mesenchymal transitions. *Nature reviews. Molecular cell biology* 7, 131-142.

Thisse, B., el Messal, M. and Perrin-Schmitt, F. (1987). The twist gene: isolation of a Drosophila zygotic gene necessary for the establishment of dorsoventral pattern. *Nucleic acids research* 15, 3439-3453.

Tjeertes, J. V., Miller, K. M. and Jackson, S. P. (2009). Screen for DNA-damage-responsive histone modifications identifies H3K9Ac and H3K56Ac in human cells. *The EMBO journal* 28, 1878-1889.

Towb, P., Bergmann, A. and Wasserman, S. A. (2001). The protein kinase Pelle mediates feedback regulation in the Drosophila Toll signaling pathway. *Development (Cambridge, England)* 128, 4729-4736.

Towb, P., Galindo, R. L. and Wasserman, S. A. (1998). Recruitment of Tube and Pelle to signaling sites at the surface of the Drosophila embryo. *Development (Cambridge, England)* 125, 2443-2450.

Tsubota, T., Berndsen, C. E., Erkmann, J. A., Smith, C. L., Yang, L., Freitas, M. A., Denu, J. M. and Kaufman, P. D. (2007). Histone H3-K56 acetylation is catalyzed by histone chaperone-dependent complexes. *Molecular cell* 25, 703-712.

Tsukada, Y., Fang, J., Erdjument-Bromage, H., Warren, M. E., Borchers, C. H., Tempst, P. and Zhang, Y. (2006). Histone demethylation by a family of JmjC domain-containing proteins. *Nature* 439, 811-816.

Vaessin, H., Caudy, M., Bier, E., Jan, L. Y. and Jan, Y. N. (1990). Role of helix-loop-helix proteins in Drosophila neurogenesis. *Cold Spring Harbor symposia on quantitative biology* 55, 239-245.

van Attikum, H., Fritsch, O., Hohn, B. and Gasser, S. M. (2004). Recruitment of the INO80 complex by H2A phosphorylation links ATP-dependent chromatin remodeling with DNA double-strand break repair. *Cell* 119, 777-788.

van der Knaap, J. A., Kumar, B. R., Moshkin, Y. M., Langenberg, K., Krijgsveld, J., Heck, A. J., Karch, F. and Verrijzer, C. P. (2005). GMP synthetase stimulates histone H2B deubiquitylation by the epigenetic silencer USP7. *Molecular cell* 17, 695-707.

van Leeuwen, F., Gafken, P. R. and Gottschling, D. E. (2002). Dot1p modulates silencing in yeast by methylation of the nucleosome core. *Cell* 109, 745-756.

Vermeulen, M., Eberl, H. C., Matarese, F., Marks, H., Denissov, S., Butter, F., Lee, K. K., Olsen, J. V., Hyman, A. A., Stunnenberg, H. G., et al. (2010). Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell* 142, 967-980.

Wade, P. A., Gegonne, A., Jones, P. L., Ballestar, E., Aubry, F. and Wolffe, A. P. (1999). Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nature genetics* 23, 62-66.

Wang, Y., Zhang, H., Chen, Y., Sun, Y., Yang, F., Yu, W., Liang, J., Sun, L., Yang, X., Shi, L., et al. (2009). LSD1 is a subunit of the NuRD complex and targets the metastasis programs in breast cancer. *Cell* 138, 660-672.

Watanabe, S., Radman-Livaja, M., Rando, O. J. and Peterson, C. L. (2013). A histone acetylation switch regulates H2A.Z deposition by the SWR-C remodeling enzyme. *Science (New York, N.Y.)* 340, 195-199.

Weake, V. M. and Workman, J. L. (2008). Histone ubiquitination: triggering gene activity. *Molecular cell* 29, 653-663.

Wei, Y., Yu, L., Bowen, J., Gorovsky, M. A. and Allis, C. D. (1999). Phosphorylation of histone H3 is required for proper chromosome condensation and segregation. *Cell* 97, 99-109.

Weingarten-Gabbay, S. and Segal, E. (2014). The grammar of transcriptional regulation. *Human genetics* 133, 701-711.

Wozniak, G. G. and Strahl, B. D. (2014). Catalysis-dependent stabilization of Bre1 finetunes histone H2B ubiquitylation to regulate gene transcription. *Genes & development* 28, 1647-1652.

Wysocka, J., Swigut, T., Xiao, H., Milne, T. A., Kwon, S. Y., Landry, J., Kauer, M., Tackett, A. J., Chait, B. T., Badenhorst, P., et al. (2006). A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature* 442, 86-90.

Xhemalce, B., Miller, K. M., Driscoll, R., Masumoto, H., Jackson, S. P., Kouzarides, T., Verreault, A. and Arcangioli, B. (2007). Regulation of histone H3 lysine 56 acetylation in Schizosaccharomyces pombe. *The Journal of biological chemistry* 282, 15040-15047.

Xu, F., Zhang, K. and Grunstein, M. (2005). Acetylation in histone H3 globular domain regulates gene expression in yeast. *Cell* 121, 375-385.

Yamane, K., Toumazou, C., Tsukada, Y., Erdjument-Bromage, H., Tempst, P., Wong, J. and Zhang, Y. (2006). JHDM2A, a JmjC-containing H3K9 demethylase, facilitates transcription activation by androgen receptor. *Cell* 125, 483-495.

Yan, J., Zierath, J. R. and Barrès, R. (2011). Evidence for non-CpG methylation in mammals. *Experimental cell research* 317, 2555-2561.

Yang, D.-J., Chung, J.-Y., Lee, S.-J., Park, S.-Y., Pyo, J.-H., Ha, N.-C., Yoo, M.-A. and Park, B.-J. (2010). Slug, mammalian homologue gene of Drosophila escargot, promotes neuronal-differentiation through suppression of HEB/daughterless.

Cell cycle (Georgetown, Tex.) 9, 2861-2874.

Yang, J., Mani, S. A., Donaher, J. L., Ramaswamy, S., Itzykson, R. A., Come, C., Savagner, P., Gitelman, I., Richardson, A. and Weinberg, R. A. (2004). Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *Cell* 117, 927-939.

Yin, Z. and Frasch, M. (1998). Regulation and function of tinman during dorsal mesoderm induction and heart specification in Drosophila. *Developmental genetics* 22, 187-200.

Yin, Z., Xu, X. L. and Frasch, M. (1997). Regulation of the twist target gene tinman by modular cis-regulatory elements during early mesoderm development. *Development (Cambridge, England)* 124, 4971-4982.

Yuan, J., Pu, M., Zhang, Z. and Lou, Z. (2009). Histone H3-K56 acetylation is important for genomic stability in mammals. *Cell cycle (Georgetown, Tex.)* 8, 1747-1753.

Zeitlinger, J., Zinzen, R. P., Stark, A., Kellis, M., Zhang, H., Young, R. A. and Levine, M. (2007). Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes & development* 21, 385-390.

Zentner, G. E. and Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. *Nature structural & molecular biology* 20, 259-266.

Zinzen, R. P., Senger, K., Levine, M. and Papatsenko, D. (2006). Computational models for neurogenic gene expression in the Drosophila embryo. *Current biology : CB* 16, 1358-1365.

Chapter 2

Characterization of lysine 56 of histone H3 as an acetylation site in *Saccharomyces cerevisiae*.

Anil Ozdemir, Salvatore Spicuglia, Edwin Lasonder, Michiel Vermeulen, Coen Campsteijn, Hendrik G. Stunnenberg and Colin Logie.

Accelerated Publication

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 280, No. 28, Issue of July 15, pp. 25949–25952, 2005

Characterization of Lysine 56 of Histone H3 as an Acetylation Site in *Saccharomyces cerevisiae**S

Received for publication, April 25, 2005,
and in revised form, May 10, 2005
Published, JBC Papers in Press, May 10, 2005,
DOI 10.1074/jbc.C500181200

Anil Ozdemir, Salvatore Spicuglia, Edwin Lasonder, Michiel Vermeulen, Coen Campsteijn, Hendrik G. Stunnenberg, and Colin Logie‡

From the Department of Molecular Biology, Nijmegen Center for Molecular Life Sciences, Radboud University, 6500 HB Nijmegen, The Netherlands

Post-translational histone modifications abound and regulate multiple nuclear processes. Most modifications are targeted to the amino-terminal domains of histones. Here we report the identification and characterization of acetylation of lysine 56 within the core domain of histone H3. In the crystal structure of the nucleosome, lysine 56 contacts DNA. Phenotypic analysis suggests that lysine 56 is critical for histone function and that it modulates formamide resistance, ultraviolet radiation sensitivity, and sensitivity to hydroxyurea. We show that the acetylated form of histone H3 lysine 56 (H3-K56) is present during interphase, metaphase, and S phase. Finally, reverse genetic analysis indicates that none of the known histone acetyltransferases is solely responsible for H3-K56 acetylation in *Saccharomyces cerevisiae*.

In eukaryotes, genetic information is packed in a higher order structure of histones and genomic DNA that is called chromatin. The fundamental unit of chromatin is the nucleosome and consists of 147 bp of DNA wrapped about twice around a histone octamer that contain a histone H3/H4 tetramer and two H2A/H2B dimers (1, 2). Post-translational modifications of the histone tails are linked to different states of chromatin that regulate processes like transcription, DNA repair, replication, and recombination (3–5). Overlapping actions of histone modifying enzymes on the very same or different histone residues generates a combinational complexity of modifications that is called the histone code (5). Hyperacetylation of lysines located in the amino-terminal tail of core histones correlates with transcriptional activation whereas hypoacetylation relates to transcriptional repression (3, 4). Histone acetylation is a dynamic process that is regulated by the opposing activities of histone acetyltransferases (HATs)[1] (6) and histone deacetylases (7). Metylation status of lysines in the amino-terminal tail, and the histone-fold domain of histone H3 plays an important role in the establishment of the active (and/or silenced) state of chromatin (5, 8).

In contrast, not much is known about histone core domain modifications and their functions. Recently, acetylation of histone H4 lysine 91 was shown to be important for chromatin assembly (9). It is also known that methylation of histone H3 lysine 79 impinges on transcription silencing (10, 11). Furthermore, a globular domain histone mutation, H3 leucine 61 to tryptophan, impaired association of SWI/SNF with chromatin (12). Here we identify and characterize acetylation of histone H3 lysine 56 as a novel core domain histone modification in *S. cerevisiae*.

MATERIALS AND METHODS

Yeast Strains, Plasmids, and Media—A list of the strains we employed is provided as supplemental Table 1. Plasmid [pHHT2-HIS3] was made by insertion of a 1010-base pair HindIII-SnaBI DNA fragment excised from [pMR366-URA3-HHT2] (13), encompassing the *HHT2* open reading frame plus 408 base pairs upstream and 210 base pairs downstream DNA. Site-directed mutagenesis on [pHHT2-HIS3] was confirmed by sequencing the entire gene. Where indicated, compounds were added to the following final concentrations; 0.2% (w/v) 5-fluoroorotic acid (5-FOA; ICN Biochemicals), 100 mM hydroxyurea (HU; Sigma), 0.01% (v/v) methyl methanesulfonate (Acros Organics), 3% (v/v) formamide (Fluka Biochemica), 15 μg/ml nocodazole (Sigma). A Stratagene UV Stratalinker was used to score sensitivity to UV irradiation.

Antiserum against Acetylated Histone H3 lysine 56 (H3-K56)—A polyclonal H3-K56[Ac] serum was raised by immunizing a rabbit with the RRFQK[Ac]STELLIRKL synthetic peptide conjugated to keyhole limpet hemocyanin.

Histone Purification—Histones were purified according to Edmondson *et al.* (14) except that zymolyase (Seikagaku Corp. catalog no. 120493) was used at a final concentration of 0.1 mg/ml.

SDS-PAGE and Western Blots—SDS-PAGE and Western blot analysis were performed according to standard procedures (15). Purified histones were separated on 15% SDS-PAGE gels and transferred to polyvinylidene difluoride membranes (Schleicher & Schuell). Membranes were incubated at 4 °C for 3 h in TBST (20 mM Tris, pH 8.0, 125 mM NaCl, and 0.05% Tween 20) with antibodies either against acetylated histone H3-K56 (1:300 dilution in TBST), diacetyl histone H3 (Upstate Biotechnology catalog number 06-599, 1:1000), acetylated histone H3-K18 (Abcam catalog number ab1191, 1:1000), tetra-acetyl histone H4 (Upstate Biotechnology catalog number 06-866, 1:1000), dimethyl histone H3-K4 (Abcam catalog number ab7766, 1:1000), trimethyl histone H3-K4 (Abcam catalog number ab8580, 1:1000), or against histone H3 (Abcam catalog number ab1791, 1:1000). Western blots were developed with an ECL detection kit (Amersham Biosciences).

Flow Cytometry Analysis—Cellular DNA content was determined as described (16) using 1 μM sytox green (Molecular Probes) and a BD Biosciences calibur fluorescence activated cell sorter.

Purification of Active HAT Fractions—Histone acetyltransferase activity was purified as described previously (17). Whole-cell extract that was prepared from a 10-liter yeast culture was loaded onto Ni^{2+}-nitrilotriacetic acid-agarose (Qiagen), eluted with 0.3 M imidazole buffer, and then applied to a Mono Q column (Amersham Biosciences). H3-K56 HAT activity eluted at 200 mM NaCl.

RESULTS AND DISCUSSION

Identification of a Novel Histone Modification—Acetylation of H3-K56, a novel core domain histone H3 modification in *S. cerevisiae* was identified multiple times by mass spectrometry analysis of histone preparations (data not shown). Zhang *et al.* (18) did not detect acetylation of histone H3 lysine 56 using calf thymus histones, although evidence for methylation

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

S The on-line version of this article (available at http://www.jbc.org) contains supplemental Tables 1 and 2.

‡ To whom correspondence should be addressed: NCMLS, P. O. Box 9101, 6500 HB Nijmegen, The Netherlands. Tel.: 31-24-3610525; Fax: 31-24-3610520; E-mail: c.logie@ncmls.ru.nl.

[1] The abbreviations used are: HAT, histone acetyltransferase; 5-FOA, 5-fluoroorotic acid; HU, hydroxyurea; H3-K56, histone H3 lysine 56; YEPD, yeast extract-peptone-dextrose.

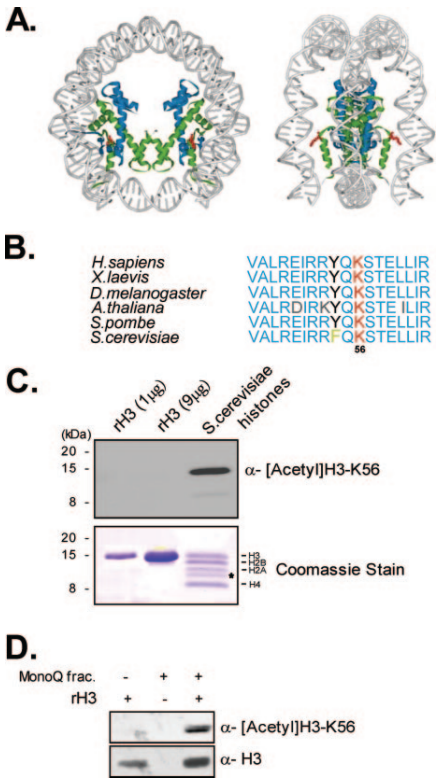


FIG. 1. **Characterization of H3-K56 acetylation.** *A*, crystal structure of the yeast nucleosome. For simplicity histone H2A/H2B dimers are not depicted. DNA is shown in *gray*, histone H3 in *green*, and H4 in *blue*. H3-K56 is highlighted in *red*. The structure is based on Protein Data Bank code 1ID3. *B*, alignment of histone H3 (amino acids 46–63) from different species. *C*, analysis of recombinant and native yeast histones. In the *upper panel*, *E. coli* expressed (*lanes 1* and *2*) and acid-extacted yeast (YN1037) histones (*lane 3*) were analyzed by Western blot using rabbit serum raised against a synthetic acetylated H3-K56 peptide. In the *lower panel*, Coomassie Blue-stained 15% SDS-polyacrylamide gel shows the quality of histone protein preparations. The protein marked with the *asterisk* is a proteolytic fragment of histone H3. *D*, a representative *in vitro* HAT assay is shown; reactions were analyzed by Western blot using an antibody against acetylated H3-K56 (*upper* panel) or the core domain of H3 (*lower* panel).

of arginine 52 or 53, or of lysine 56, of histone H3 was obtained. Crystal structure analysis of the nucleosome (1) revealed that H3-K56 is located on a side of the H3/H4 tetramer facing DNA (Fig. 1*A*). To characterize this new modification, we raised an antibody against a synthetic peptide carrying acetylated H3-K56. The antibody recognizes a protein band that comigrates with purified histone H3 (Fig. 1*C*, *lane 3*). It does not recognize recombinant yeast histone H3 expressed in *Escherichia coli* (rH3) (Fig. 1*C*, *lanes 1* and *2*). However, the antiserum does recognize rH3 after an *in vitro* HAT reaction using active yeast extract (Fig. 1*D*, *lane 3*). Phenylalanine at position 54 of histone H3 is not conserved in other species (Fig. 1*B*), therefore the antiserum is specific for the acetylated form of *S. cerevisiae* histone H3-K56, and it does not recognize mammalian histone H3 (data not shown).

Analysis of the Mutant hht2 Alleles—To further validate that the antibody specifically recognizes acetylated H3-K56, and to gain insight into the possible function(s) of this modification, we constructed budding yeast strains that expressed wild type and mutant alleles of *HHT2* from low copy number plasmids as sole source of histone H3. The effect on viability of point mutations at position 56 was assayed in a yeast strain, YN1375, lacking both chromosomal copies of H3. This strain harbored wild type *HHT2* on a *URA3* plasmid. Hence, medium containing 5-FOA did not permit growth of YN1375 (Fig. 2*A*). The single amino acid substitutions of histone H3 lysine 56 to alanine (H3-K56A) or to arginine (H3-K56R) borne by the *HIS3* plasmid sustained viability of YN1375 on 5-FOA plates, indicating functionality (Fig. 2*A*). In contrast histone H3 bearing a glutamate at position 56 (H3-K56E) could not support cell proliferation (Fig. 2*A*).

Both H3-K56A and H3-K56R substitutions disrupted the epitope as such that H3-K56 [acetyl] antibody recognition of the mutant H3 histones was abolished (Fig. 2*B*, *panel 2*). To exclude the possibility that the level of histone H3 was affected in the mutants we used a commercial antibody that recognizes another epitope within the core domain of histone H3. As shown in Fig. 2*B* (*panel 1*), the total amount of histone H3 is similar in all strains. These results indicate that the antiserum we raised is highly specific for acetylated H3-K56.

Interplay with Other Histone Modifications—A particular modification that is present on a histone residue may coexists with, or be required for, modifications at other residues (3). Acetylation of lysines that are located at the N-terminal tail of histones H3 and H4 are associated with transcription activation (4, 5). We sought to find out whether the acetylation of H3-K56 was a determinant of known histone tail modifications. To this end, we purified histones from strains expressing H3-K56A (YN1392) or H3-K56R (YN1393) as a sole source of histone H3. Global acetylation levels of histone H3 and histone H4 N-terminal tails were not affected in the *hht2-K56A* and *hht2-K56R* mutants (Fig. 2*B*, *panels 3–5*). The levels of di- and trimetylation of histone H3-K4 were not different either (Fig. 2*A*, *panels 6* and *7*). These findings suggest that H3-K56 acetylation is not required for the establishment and/or the maintenance of these epigenetic marks at the genome wide level. We note that this does not exclude the possibility that acetylation of H3-K56 might influence the levels of histone modifications at specific loci.

Phenotype Analysis of the hht2-K56A and hht2-K56R Alleles—To better understand the function of H3-K56 acetylation, we performed a phenotypic analysis on the *hht2-K56A* and *hht2-K56R* alleles. Single amino acid substitution of a lysine to an arginine (*hht2-K56R*) is predicted to cause no major changes within the structure of the H3/H4 tetramer. Because of the position of the residue (Fig. 1*A*); however, we expect to retain ionic interactions between histone H3 and DNA, which would promote a more stable chromatin template. Alanine on the other hand is a smaller amino acid than lysine and is not charged. Therefore substitution to an alanine (*hht2-K56A*) is expected to weaken the interactions between histone H3 and DNA, thereby destabilizing the nucleosome and creating a more flexible environment for chromatin remodelers and transcription associated regulatory protein complexes.

Temperature sensitivity is a common yeast phenotype (19). Surprisingly, the *hht2-K56A* allele conferred a growth advantage to the cells at 37 °C relative to the *HHT2* and the *hht2-K56R* alleles (Fig. 2*C*, *first* row). It has been reported that 30% of formamide-sensitive strains also display temperature sensitivity (19). We therefore also tested formamide sensitivity. Not much is known about the molecular mechanisms that underpin

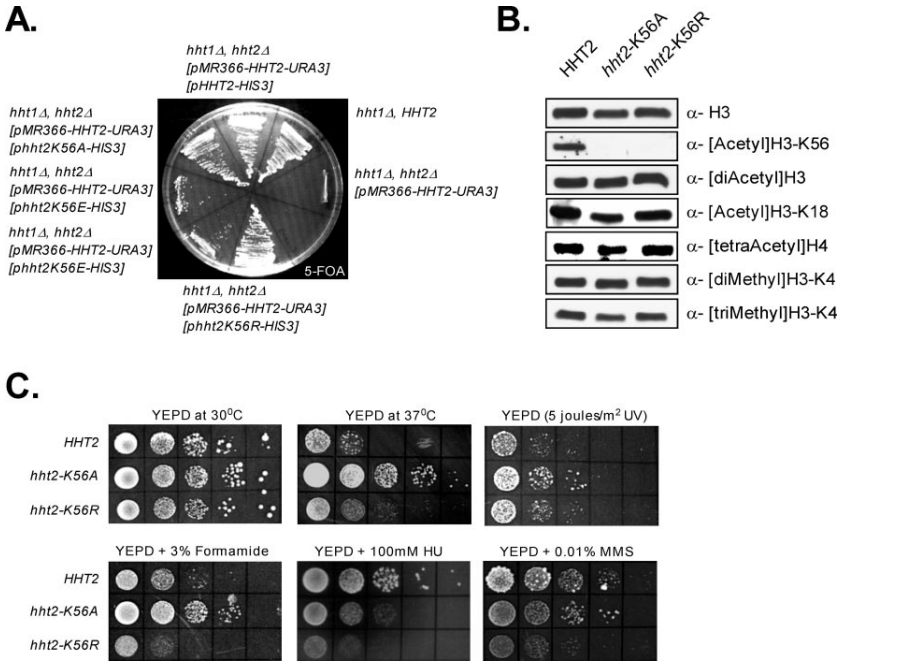


FIG. 2. **Analysis of histone H3 lysine 56 point mutations.** *A*, the *hht2-K56A* and *hht2-K56R* alleles support viability, *hht2-K56E* does not. *B*, analysis of other histone H3 modifications in H3-K56 mutants. Histones from the stains carrying either the wild type *HHT2* (YN1391, *lane 1*), or the mutant *hht2-K56A* (YN1392, *lane 2*), and *hht2-K56R* (YN1393, *lane 3*) alleles of histone H3 were analyzed by Western blot using antibodies against a core domain histone H3 (*panel 1*), acetylated H3-K56 (*panel 2*), diacetyl histone H3 (*panel 3*), acetyl H3-K18 (*panel 4*), tetra-acetyl histone H4 (*panel 5*) and against di- and trimethyl histone H3-K4 (*panels 6* and *7*). *C*, phenotypes conferred by the *hht2-K56A* and *hht2-K56R* alleles. Cells were grown to mid-log phase at 30 °C in selective SD media, and then 10-fold serial dilutions were spotted on the indicated medium. Pictures of plates incubated at 30 °C for 3–4 days are shown.

this phenotype, although it likely reflects hydrogen bridge destabilization. The *hht2-K56A* allele also displayed a growth advantage on YEPD containing 3% formamide (Fig. 2*C*, *second* row), whereas yeast strains carrying either the *wild type* or the *hht2-K56R* alleles of histone H3 were clearly defective for growth on this medium. This would suggest that suppression of the lethality induced by formamide is not a result of the loss of acetylation at lysine 56 but that it is associated with a structural advantage conferred by the alanine substitution onto the nucleosome.

HU is an inhibitor of ribonucleotide reductase; hence exposure to HU causes yeast cells to arrest in S phase of the cell cycle. Growth of both mutant strains was clearly retarded on YEPD + HU relative to the wild type strain, and the effect of the *hht2-K56R* allele was much more pronounced (Fig. 2*C*, *second* row). The same results were obtained when methyl methanesulfonate was used instead of HU (Refs. 16 and 19; Fig. 2*C*, *second* row). This phenotype implies a possible role of H3-K56 acetylation in DNA replication-coupled repair and/or progression through the S phase of the cell cycle.

Sensitivity to UV irradiation indicates defects in DNA damage repair responses. Mutants bearing either *hht2-K56A* or *hht2-K56R* alleles of histone H3 showed a significant increase of survival when exposed to 5 joules/m^2 of UV irradiation (Fig. 2*C*, *first* row). We envisage two explanations for this phenotype; either a lethal DNA damage-induced cell cycle block is circumvented, or the repair pathway is constitutively on in the

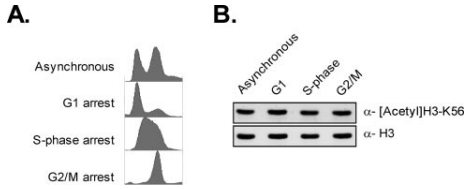


FIG. 3. **Acetylation of H3-K56 during the cell cycle.** *A*, flow cytometry analysis. *Wild type* cells were used for asynchronous cultures (YN1037, *panel 1*). *cdc25-2* cells were arrested in G_1 by 5 h of heat shock at 37 °C (YN133, *panel 2*). S phase and G_2/M arrests were achieved by growing yeast (YN1037) for 4 h in YEPD containing hydroxyurea or nocodazole, respectively (*panels 3* and *4*). *B*, histones extracted from cell cycle staged yeast (Fig. 3A) were analyzed by Western blot using antiserum against acetylated H3-K56 (*panel 1*). *Lanes 1–4* correspond to histones purified from asynchronous, G_1-, S-phase, and G_2/M-arrested cells, respectively. Relative amount of the histone H3 in each sample was quantified using an antibody against histone H3 (*panel 2*).

mutant strains. This could be due to a direct involvement of H3-K56 acetylation in repair process or indirectly via an altered cellular transcription related profile.

Cell Cycle Regulation of H3-K56 Acetylation—The fact that the mutant *hht2-K56A* and *hht2-K56R* alleles of histone H3 showed DNA damage repair and replication-related phenotypes may be taken to indicate that H3-K56 acetylation takes

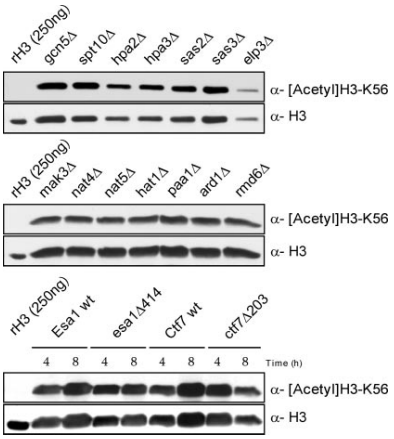


FIG. 4. **Screen to identify the HAT responsible for acetylation of H3-K56 (see supplemental Table 1 for strains).** Genetic deletion mutants for non-essential HATs (supplemental Table 2) were grown to saturation at 30 °C in YEPD. Cells carrying either wild type or temperature-sensitive (*ts*) alleles of the essential *Esa1* and *Ctf7* acetyltransferases were grown to mid-log phase at 25 °C, then shifted to 37 °C. Samples were collected after 4 and 8 h of heat shock. Histones were extracted from all samples and analyzed for the presence of the modification by Western blot.

place at a defined stage of the cell cycle. To examine this possibility, we assayed for the presence of H3-K56 acetylation in G_1-, S-, and G_2/M phase-arrested *S. cerevisiae* (Fig. 3*A*). This revealed that H3-K56 acetylation is present in G_1, S-phase, and G_2/M (Fig. 3*B*).

Screen for the H3-K56 Acetyltransferase—To identify the HAT responsible for this novel histone modification, we performed a screen with deletion strains of the major putative HATs (supplemental Table 1). There are two classes of HATs: the A-type HATs are located in the nucleus and acetylate nucleosomal histones, and the B-type HATs on the other hand are located in the cytoplasm and acetylate free histones (6). Because H3-K56 is likely to bind DNA (Fig. 1*A*), we expect that the acetylation occurs on free histone H3. To be accurate, however, we included both classes in our experiments (supplemental Table 2). Esa1p and Ctf7p are essential acetyltransferases (20, 21). For this reason we used strains that express temperature-sensitive alleles of *ESA1* and *CTF7* (supplemental Tables 1 and 2). The results presented in Fig. 4 show that all the HAT deletion strains and the cells harboring mutant alleles of *ESA1* and *CTF7* retained the H3-K56 acetylation. This suggests either that an as yet unidentified HAT exists or that multiple HATs can acetylate H3-K56.

The identification and genetic characterization of H3-K56 acetylation suggests physiological roles for this histone modification in *S. cerevisiae*. Reversal of the charge at this position (H3-K56E) is lethal (Fig. 2*A*). This indicates that lysine 56 plays a pivotal role in chromatin structure. The fact that this residue is acetylated underscores the notion that histone core domain residues have biological functions that extend beyond a simple structural role and contribute to regulate chromatin remodeling (22).

Acknowledgments—We thank Karolyn Luger for rH3 protein, Hans Adams for peptide synthesis, Xavier Le Guezennec for help with molecular modeling, Jacques Cote and Robert Skibbens for yeast strains, and Craig L. Peterson for the [pMR366-HHT2] plasmid.

REFERENCES

1. White, C. L., Suto, R. K., and Luger, K. (2001) *EMBO J.* **20,** 5207–5218
2. Luger, K. (2003) *Curr. Opin. Genet. Dev.* **13,** 127–135
3. Margueron, R., Trojer, P., and Reinberg, D. (2005) *Curr. Opin. Genet. Dev.* **15,** 163–176
4. Narlikar, G. J., Fan, H. Y., and Kingston, R. E. (2002) *Cell* **108,** 475–487
5. Jenuwein, T., and Allis, C. D. (2001) *Science* **293,** 1074–1080
6. Sterner, D. E., and Berger, S. L. (2000) *Microbiol. Mol. Biol. Rev.* **64,** 435–459
7. Ng, H. H., and Bird, A. (2000) *Trends Biochem. Sci.* **25,** 121–126
8. Lachner, M., and Jenuwein, T. (2002) *Curr. Opin. Cell Biol.* **14,** 286–298
9. Ye, J., Ai, X., Eugeni, E. E., Zhang, L., Carpenter, L. R., Jelinek, M. A., Freitas, M. A., and Parthun, M. R. (2005) *Mol. Cell* **18,** 123–130
10. van Leeuwen, F., Gafken, P. R., and Gottschling, D. E. (2002) *Cell* **109,** 745–756
11. Ng, H. H., Ciccone, D. N., Morshead, K. B., Oettinger, M. A., and Struhl, K. (2002) *Genes Dev.* **16,** 1518–1527
12. Duina, A. A., and Winston, F. (2004) *Mol. Cell. Biol.* **24,** 561–572
13. Kruger, W., Peterson, C. L., Sil, A., Coburn, C., Arents, G., Moudrianakis, E. N., and Herskowitz, I. (1995) *Genes Dev.* **9,** 2770–2779
14. Edmondson, D. G., Smith, M. M., and Roth, S. Y. (1996) *Genes Dev.* **10,** 1247–1259
15. Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., and Struhl, K. (2003) *Current Protocols in Molecular Biology*, Wiley Interscience, Hoboken, NJ
16. Foss, E. J. (2001) *Genetics* **157,** 567–577
17. Eberharter, A., John, S., Grant, P. A., Utley, R. T., and Workman, J. L. (1998) *Methods* (*Orlando*) **15,** 315–321
18. Zhang, L., Eugeni, E. E., Parthun, M. R., and Freitas, M. A. (2003) *Chromosoma* (*Berl.*) **112,** 77–86
19. Hampsey, M. (1997) *Yeast* **13,** 1099–1133
20. Skibbens, R. V., Laura, B., Corson, L. B., Koshland, D., and Hieter, P. (1999) *Genes Dev.* **13,** 307–319
21. Eisen, A., Utley, R. T., Nourani, A., Allard, S., Schmidt, P., Lane, W. S., Lucchesi, J. C., and Cote, J. (2001) *J. Biol. Chem.* **276,** 3484–3491
22. Xu, F., Zhang, K., and Grunstein, M. (2005) *Cell* **121,** 375–385

Chapter 3

Histone H3 lysine 56 acetylation: a new twist in the chromosome cycle.

Ozdemir A, Masumoto H, Fitzjohn P, Verreault A, and Logie C.

[Cell Cycle 5:22, 2602-2608, 15 November 2006]; ©2006 Landes Bioscience

Review

Histone H3 Lysine 56 Acetylation

A New Twist in the Chromosome Cycle

Anil Ozdemir[1]
Hiroshi Masumoto[2]
Paul Fitzjohn[3]
Alain Verreault[4]
Colin Logie[1,*]

[1]Department of Molecular Biology; Nijmegen Center for Molecular Life Sciences; Radboud University; The Netherlands

[2]Laboratories for Biomolecular Networks; Graduate School of Frontier Biosciences; Osaka University; Suita, Osaka Japan

[3]Cancer Research UK; London Research Institute; London UK

[4]Institut de Recherche en Immunologie et Cancérologie; Université de Montréal; Montréal, Québec Canada

*Correspondence to: Colin Logie; Department of Molecular Biology; Nijmegen Center for Molecular Life Sciences; Radboud University; 6500 HB Nijmegen The Netherlands; Tel.: +31.24.3610525; Fax: +31.24.3610520; Email: c.logie@ncmls.ru.nl

Original manuscript submitted: 09/12/06
Revised manuscript submitted: 10/02/06
Manuscript accepted: 10/02/06

Previously published online as a *Cell Cycle* E-publication:
http://www.landesbioscience.com/journals/cc/abstract.php?id=3473

KEY WORDS

histone, acetylation, nucleosome, DNA damage, cell cycle, chromatin

ACKNOWLEDGEMENTS

We thank A. Salcedo and H. Stunnenberg for sharing unpublished data. Research in A. Verreault's laboratory is funded by the Canadian Institutes for Health Research (CIHR). Research in C. Logie's Laboratory is funded by KWF and EuroDYNA. H. Matsumoto is grateful to Professor Akio Sugino for laboratory space and reagents.

ABSTRACT

Several recent reports have identified lysine 56 (K56) as a novel site of acetylation in yeast histone H3. K56 acetylation is predicted to disrupt some of the histone-DNA interactions at the entry and exit points of the nucleosome core particle. This modification occurs in virtually all the newly synthesised histones that are deposited into chromatin during S-phase. Cells with mutations that block K56 acetylation show increased genome instability and hypersensitivity to genotoxic agents that interfere with replication. Removal of K56 acetylation takes place in the G_2/M phase of the cell cycle and is dependent upon Hst3 and Hst4, two proteins that are related to the NAD^+-dependent histone deacetylase Sir2. In response to DNA damage checkpoint activation during S-phase, expression of Hst3/Hst4 is delayed to extend the window of opportunity in which K56 acetylation can act in the DNA damage response. The high abundance of histone H3 K56 acetylation, its regulation and strategic location in the nucleosome core particle raise a number of fascinating issues that we discuss here.

INTRODUCTION

Core histones consist of two domains: an N-terminal tail and a globular domain. Until a few years ago, histone post-translational modifications were largely confined to the N-terminal tails, which protrude beyond the DNA gyres and are therefore relatively accessible to histone modifying enzymes. Histone modifications were originally identified by N-terminal Edman sequencing of intact histone proteins, a technique that was not sufficiently sensitive to identify modifications that occurred far beyond the N-terminal tails in the primary amino acid sequence. In recent years, this limitation has been largely overcome through the advent of mass spectrometry. This has led to the discovery of a myriad of novel histone modifications, many of which have yet to be ascribed a biological function.[1,2] Six independent research groups recently reported the discovery of lysine 56 as a novel site of histone H3 acetylation in the budding yeast *Saccharomyces cerevisiae*.[3-8] K56 acetylation has also been observed in the fission yeast *Schizosaccharomyces pombe*,[6] which is evolutionarily very distant from *S. cerevisiae*.[9] Based on mass spectrometry, K56 acetylation occurs in *Plasmodium falciparum* (A. Salcedo and H. Stunnenberg, *in preparation*) and the modification has also been reported in *Drosophila*.[7] However, K56 acetylation has thus far not been detected in mammalian cells.[5,7]

Lysine 56 (K56) is the last residue of an α-helix, known as αN, that connects the N-terminal tail to the globular domain of H3 (Fig. 1). There are two symmetry-related H3 molecules in the nucleosome core particle and the positive charges of the K56 ε-amino group in each H3 molecule make water-mediated contacts with DNA segments near the entry and exit points in the nucleosome (Fig. 1). In addition to K56, other residues of the αN helix also contact DNA at the same sites in the nucleosome.[10] Thus, although K56 acetylation likely weakens these contacts, it is probably not sufficient to disrupt them completely. In *S. cerevisiae*, R52 is essential for viability and cannot be mutated into alanine, lysine or glutamine.[3] Interestingly, based on mass spectrometry, R52, R53 or K56 is monomethylated in bovine histones.[2] As there was no peptide fragmentation in this analysis, it was not possible to assign the mono-methylation to a specific residue. Nevertheless, this result argues that vertebrates also modulate this portion of the nucleosome core through post-translational modification. A further complication is that terminal DNA segments in the nucleosome rapidly dissociate and rebind to the histone surface.[11,12] This dynamic equilibrium of the DNA with respect to the underlying histone surface occurs spontaneously even in the absence of H3 K56 acetylation.[11,12] This is generally consistent

with the fact that histones make fewer contacts with DNA at the entry and exit points than at any other site in the nucleosome.[10] This may facilitate the recognition of K56-acetylated histone H3 by deacetylases and possibly other nucleosome remodeling enzymes. In vivo, cells that lack K56 acetylation exhibit a chromatin structure where the DNA is more extensively supercoiled and less accessible to nucleases than wild-type cells.[13,14] However, it is not clear whether this global relaxation of chromatin structure is directly achieved by K56 acetylation or through its recognition by proteins that elicit perturbations in nucleosome structure that are more extensive than what is possible with K56 acetylation alone. Yeast cells where the only available source of histone H3 cannot be acetylated at K56 are viable. For instance, K56 can be substituted by non-acetylatable residues, such as alanine or arginine.[3-8] Thus, the modification is not essential for cell viability or de novo nucleosome assembly. However, the introduction of a negatively charged glutamate residue at position 56 is lethal unless wild type histone H3 is also present in the same cells.[5] The reason why the H3 K56E mutation is lethal is not known. Nonetheless, this result reinforces the notion that lysine 56 plays a pivotal role in chromatin function.

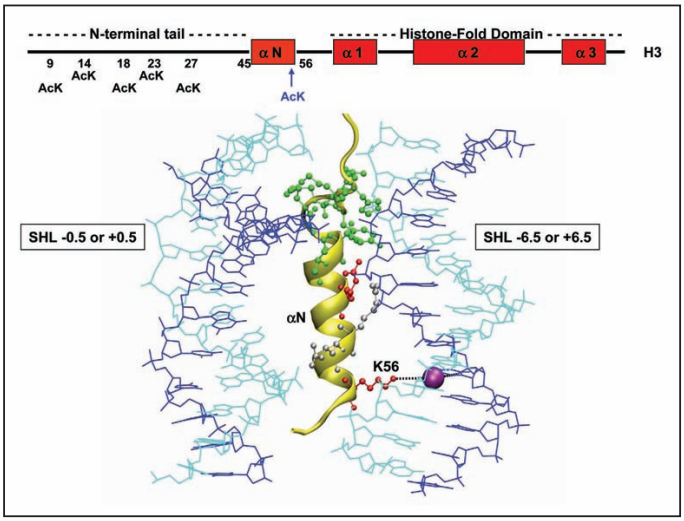


Figure 1. The αN helix of H3 is located between DNA duplexes at the entry/exit point (right) and in the middle of the nucleosome (left). The K56 side chain interacts with the DNA via a water molecule (mauve sphere). Residues 39-46 of H3 (green) make alternating contacts with the two DNA duplexes.

H3 K56 ACETYLATION: WHERE AND WHEN?

Some controversy had arisen as to exactly when K56 acetylation was removed during the cell cycle. Using *MATa* cells released from a G_1 arrest with α-factor, a number of groups reported that K56 acetylation increases during S-phase.[4,6,8,15] This is because the bulk of H3 K56 acetylation occurs in newly synthesised histones that are deposited in the wake of DNA replication forks during S-phase.[4,8] H3 K56 acetylation also occurs during premeiotic S-phase and is needed for meiosis.[6] However, H3 synthesized outside of S-phase can be acetylated on K56, showing that K56 acetylase activity is not restricted to S-phase.[4] Using cell division cycle (*cdc*) mutants to arrest yeast cells at different stages of the cell cycle, high levels of K56 acetylation were also detected outside of S-phase.[5] All the studies of K56 acetylation thus far were performed with five distinct affinity-purified polyclonal antibodies raised against synthetic peptides.[4-8] The specificity of all these antibodies for K56 acetylation was rigorously demonstrated by the absence of signal in yeast strains where histone H3 K56 was mutated into a nonacetylatable residue. However, the presence of a number of potentially modified residues close to H3 K56, including R52 and R53 (Fig. 2A), raised the possibility that some antibodies may be influenced by cell cycle-modulated second-site modifications near H3 K56.

To examine this possibility, two distinct antibodies were used to probe the same cell extracts. In previous studies, one of these antibodies was able to detect high levels of K56 acetylation outside of S-phase,[5] whereas the other one was not.[4] Here we show that both antibodies essentially generate the same results. Notably, both antibodies show substantially less K56 acetylation in G_1 cells arrested with α-factor than cells released from α-factor into an hydroxyurea (HU) arrest (Fig. 2). HU blocks DNA replication by depleting pools of deoxyribonucleoside triphosphates.[16] Thus, both antibodies confirmed that levels of K56 acetylation rise during S-phase. Surprisingly, when cells were arrested either in G_1 or G_2/M by thermosensitive *cdc* mutations, both antibodies detected high levels of K56 acetylation (Fig. 2). G_1 arrest was achieved with a *ts* mutation in the yeast Ras1 and Ras2 guanine nucleotide exchange factor Cdc25.[17] It therefore appears that the acetylation status of H3 K56 is differentially modulated when cells are blocked in G_1 by triggering the pheromone response pathway or by interfering with the Ras dependent cyclic AMP signaling pathway that mediates metabolic control. G_2/M arrest was triggered by galactose-inducible expression of Swe1, which inhibits Cdk1 by phosphorylation of tyrosine 19 in Cdk1.[18] This arrest resulted in persistence of K56 acetylation (Fig. 2).[5] The deacetylation of H3 K56 depends on Hst3 and Hst4, two proteins that belong to the Sir2 family of NAD^+-dependent deacetylases.[15,19] The *HST3* gene is a member of the *CLB2* cluster of mRNAs that are expressed concomitantly with the mRNA encoding the mitotic cyclin Clb2.[20] Consistent with this, maximal expression of the Hst3 and Hst4 proteins occurs late in the cell cycle.[15] The fact that overexpression of Swe1 results in high levels of K56 acetylation in G_2/M suggests that the levels of Hst3/Hst4 proteins and/or their ability to promote K56 deacetylation may depend upon G_2/M phase CDK activity. Alternatively, cells may need to complete a CDK-dependent event before H3 K56 deacetylation can be initiated.

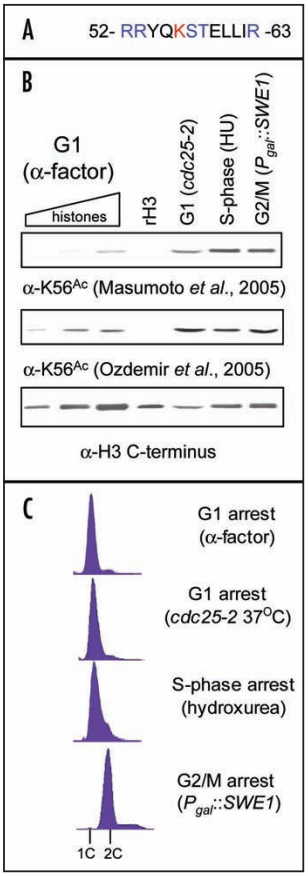


Figure 2. A) Amino acid sequence flanking K56 (red) in *S. cerevisiae* histone H3. Potentially modifiable residues are shown in blue. B) Western blot analysis of whole-cell lysates using two different anti-H3 K56Ac antibodies[4,5] and a control antibody directed against the C-terminus of H3 (Abcam ab1791). α-factor and HU-arrested cells were in the W303 background (YN2), *cdc25-2* cells were purchased from Stratagene (cat# 37.81, YN133), *Pgal::SWE1* cells were in the W303 background (YN207). Recombinant yeast H3 was a kind gift of Dr. K. Luger. C) DNA content was determined by staining with Sytox Green and flow cytometry.[84]

During DNA replication, parental histone H3/H4 located ahead of the replication fork are transferred onto both nascent sister chromatids behind the fork.[21] Conceivably, parental histones could be rapidly acetylated and deacetylated during this process. However, K56 acetylation is undetectable when cells go through S-phase in the absence of de novo histone synthesis.[4] Moreover, inhibition of Hst3/Hst4 during a single round of S-phase results in 50% K56 acetylation in G_2.[19] This implies that the vast majority of newly synthesized histones deposited throughout the genome during S-phase are K56-acetylated and that there is little or no K56 acetylation in parental histones during replication. The existence of significant K56 acetylation turnover in parental histones should have resulted in K56 acetylation rising substantially above 50% after a round of S-phase in the absence of Hst3/Hst4. This was clearly not the case.[19] The possibility that K56 acetylation in parental histones may be turned over by enzymes other than Hst3/Hst4 is unlikely because mutations of all the other known histone deacetylases do not increase K56 acetylation.[19]

H3 K56 ACETYLTRANSFERASES

K56 acetylation is reduced by about 15% in *spt10* null mutants.[7] Spt10 contains sequence motifs characteristic of acetyltransferases.[22] In addition, residues that are predicted to be required for acetyltransferase activity are indeed important for Spt10 function in vivo.[23] However, there is currently no in vitro evidence that Spt10 directly acetylates histone H3 K56 either in free or nucleosomal histones. Spt10 contains a site-specific DNA binding domain that cooperatively recognizes pairs of upstream activating sequences, known as histone UAS elements.[24,25] Pairs of these elements are present in the divergent promoters of the gene pairs encoding H2A-H2B and H3-H4, but are conspicuously absent from other yeast promoters.[24] This strongly suggests that Spt10 is a transcription factor dedicated to histone gene expression. However, neither Spt10 nor K56 acetylation are absolutely essential for histone gene expression. Although Spt10 binds to all the major core histone gene promoters in vivo, only a subset of these genes are severely crippled in their expression in the absence of Spt10.[23,24] Cells lacking Spt10 have global defects in chromatin structure and exhibit a prolonged cell cycle progression delay.[23,24] The latter phenotype is strongly suppressed by extra copies of H2A-H2B and H3-H4 genes.[24] The fact that 85% of K56 acetylation remains in *spt10* null mutants argues for the existence of other enzymes that acetylate H3 K56. Gcn5 and Hat1 have been implicated in the acetylation of newly synthesized H3 and H4.[26,27] However, Gcn5, Hat1 and several other putative or known histone acetyltransferase (HAT) catalytic subunits are dispensable for K56 acetylation in vivo.[5] Thus, the enzyme(s) responsible for the bulk of histone H3 K56 acetylation are currently unknown. It is formally possible that more than one of the currently known HATs function in a redundant manner to acetylate H3 K56. Alternatively, the K56 acetylase may belong to an as yet undefined HAT family.

CONSTITUTIVE K56 ACETYLATION RESULTS IN SPONTANEOUS DNA DAMAGE

In *hst3 hst4* double mutants, essentially the whole genome (98% of H3) is K56-acetylated even in G_1.[19] Cells lacking Hst3/Hst4 are thermosensitive but, even at the permissive temperature, they experience abnormally high levels of spontaneous DNA damage during replication.[19,28] Unlike in wild-type cells, a significant portion of the damage persists in *hst3 hst4* mutants,[19] suggesting that at least some replication-linked DNA lesions are impossible to repair in these mutants. Consistent with this, even at 25°C, *hst3 hst4* mutants contain a substantial fraction of inviable cells.[28] In addition, these mutants are exquisitely sensitive to perturbations of the replisome that are well tolerated by wild-type cells.[19,29-31] High rates of spontaneous DNA damage are only apparent in the second round of S-phase following inactivation of Hst3 in *hst4* mutants.[19] This is likely

because, during the first S-phase, K56 acetylation is confined behind replication forks. In contrast, during a second S-phase in the absence of Hst3/Hst4, K56 acetylation is present both in front of and behind replication forks. At least some of the spontaneous damage that occurs during replication likely reflects DNA double-strand breaks (DSBs) because both Rad52 and the three subunits of the MRN complex (Mre11, Rad50, Nbs1) are essential for viability in *hst3 hst4* mutants.[19] Rad52 and the MRN complex are involved in homologous recombination (HR) between sister chromatids,[32,33] which is the major pathway to repair DSBs generated during replication in *S. cerevisiae*. Based on mass spectrometry, none of the other known sites of H3 or H4 acetylation are affected in *hst3 hst4* mutants.[19] Remarkably, point mutation of H3 K56 into a non-acetylatable arginine residue suppresses the phenotypes of *hst3 hst4* mutants.[15,19] These results strongly argue that the inappropriate presence of K56 acetylation in front of replication forks results in frequent DSBs during replication. The need to avoid high levels of K56 acetylation in parental histones prior to the onset of DNA replication is probably sufficient to explain why K56 acetylation is normally removed by Hst3/Hst4 in the late stages of the cell cycle.[4,15,19] It is also formally possible that genome-wide K56 acetylation in G_2 could interfere with mitotic chromosome segregation. This notion is supported by the fact that *hst3 hst4* mutants have a high incidence of mitotic chromosome loss that is suppressed by an H3 K56R mutation.[19,28] However, the chromosome loss phenotype of *hst3 hst4* mutants could also reflect the presence of unrepairable DNA damage, rather than defects in segregation of undamaged chromosomes.

Asf1 IS ESSENTIAL FOR H3 K56 ACETYLATION

Asf1 is an evolutionarily conserved histone chaperone that was biochemically purified by virtue of its ability to enhance Chromatin Assembly Factor 1 (CAF-1)-mediated nucleosome assembly onto replicating DNA.[34] In yeast and higher eukaryotes, Asf1 functions in replication-dependent nucleosome assembly mediated by CAF-1 and a transcription-coupled nucleosome assembly pathway that depends upon Hir proteins.[35-40] Genetic studies in yeast revealed that Asf1 plays a unique role in the DNA damage response. Cell lacking Asf1 are more sensitive than *caf1* or *hir* mutants to a number of genotoxic agents that predominantly cause DNA damage by interfering with DNA replication fork progression, such as HU, camptothecin (CPT) and methyl methane sulphonate (MMS).[34,35,41] Sensitivity to these agents is also observed in H3 K56R mutant cells.[3-6] Asf1 forms a complex with Rad53 that dissociates in response to DNA damage.[42,43] Rad53 is related to human CHK1 and CHK2 and is a key protein kinase in the DNA damage response in *S. cerevisiae*.[44] Because most of the Rad53 protein is bound to Asf1,[42] it was generally assumed that the role of Asf1 in the DNA damage response would depend upon its damage-regulated interaction with Rad53. The exact role of the Asf1-Rad53 interaction in the response to genotoxic stress is not known. In contrast, two recent studies showed that Asf1 is essential for K56 acetylation in *S. cerevisiae*.[6,19] Moreover, much of the DNA damage sensitivity of *asf1* mutants can be accounted for by the loss of K56 acetylation.[6] The Asf1 protein does not contain any of the catalytic site motifs that define acetyltransferases.[22] It seemed possible that Asf1 might act by controlling Hst3/Hst4 to prevent premature deacetylation of newly synthesized histones prior to their incorporation into chromatin. However, this is not the case because the Asf1 protein is still needed for K56 acetylation in the absence of Hst3 and Hst4.[19] Thus, the available data argues that Asf1 is somehow needed for K56 acetylation of newly synthesised H3 molecules. Interestingly, Asf1 also binds to the SAS complex (Sas2, Sas4, Sas5).[45,46] This suggests that Asf1 may serve as a substrate presentation molecule to enhance the acetylation of newly synthesized histones by specific HATs. This hypothesis is consistent with the fact that Asf1 point mutations that cripple K56 acetylation and confer DNA damage sensitivity are located in a surface that mediates its interaction with histone H3.[6,47,85] However, the SAS complex is not required for K56 acetylation[5] and the binding of Asf1 to H3/H4 completely blocks the ability of the purified SAS complex to acetylate H3 and H4.[48] Thus, the mechanism by which Asf1 promotes K56 acetylation in vivo is not known.

A ROLE FOR K56 ACETYLATION IN THE RESPONSE TO REPLICATION-LINKED DNA DAMAGE

The fact that the DNA damage sensitivity of *asf1* mutants largely stems from their lack of K56 acetylation[19] provides important clues regarding the role of K56 acetylation in the response to genotoxic agents. As stated earlier, H3 K56R and *asf1* mutants are both sensitive to genotoxic agents that predominantly cause DNA DSBs by interfering with replication fork progression. In haploid yeast cells, HR between sister chromatids is the most efficient pathway to repair DSBs during S-phase and G_2.[49] Cells lacking Asf1, and therefore also devoid of K56 acetylation, do not have prominent defects in the repair of site-specific DSBs by HR.[14,50,51] The other major pathway to repair DSBs, nonhomologous DNA end-joining (NHEJ) is not substantially impaired in the absence of K56 acetylation.[4,50] In addition, cells lacking K56 acetylation are far less sensitive to ionizing radiation than HR mutants and can repair heavily fragmented chromosomes in G_2/M phase of the cell cycle.[4,51] Furthermore, K56 acetylation is not detectably induced in response to DSBs caused by ionizing radiation in G_1 or G_2/M phase (stages of the cell cycle when the abundance of K56 acetylation is rather low), even though these cells can repair DSBs very effectively.[4]

Collectively, these results argue that cells have K56 acetylation-independent mechanisms to repair DSBs that occur in the context of mature chromatin during both G_1 and G_2/M phase of the cell cycle. Consistent with this, several HATs are recruited to DSBs, such as Esa1 (TIP60 in human cells), Hat1 and Gcn5.[52-56] Several of these enzymes acetylate multiple lysine residues in the N-terminal tails of H3 and H4, but they are all individually dispensable for K56 acetylation.[5] In addition, several lysine residues in the N-terminal tails of H3 and H4 need to be mutated simultaneously to confer significant DNA damage sensitivity.[52,56,57] In contrast, even though it does not affect bulk levels of acetylation of the N-termini of H3 and H4,[5] the H3 K56R mutation alone is sufficient to confer a pronounced degree of sensitivity to clastogenic agents that result in DNA breaks during replication.[3-5] This suggests that the role of K56 acetylation is unique and may be restricted to DNA damage that arises during S-phase when chromatin regions in front of and behind replication forks are not fully mature. Interestingly, a number of abnormal perturbations of the replisome occur when replication forks are blocked in the absence of K56 acetylation. Based on chromatin immunoprecipitation assays, several replisome proteins dissociate when replication forks are stalled with HU in *asf1* mutants.[35] In contrast, DNA polymerase α aberrantly accumulates at HU-blocked forks in *asf1* mutants.[35] These events correlate with excessive uncoupling of MCM proteins from the replisome.[35] The MCMs likely act as the replicative DNA helicase.[58] When DNA lesions halt replication,

transient uncoupling of the helicase from the replisome is necessary to generate an extended stretch of single-stranded DNA.[59] This step is a prerequisite to activate DNA damage checkpoint kinases,[59] which are important to stabilize the stalled replisome and promote resumption of DNA synthesis.[60] However, extensive uncoupling of MCMs from the replisome could create inappropriately long stretches of single-stranded DNA in front of stalled forks. This may generate the substrate necessary for accumulation of DNA polymerase α when HU blocks replisome progression in *asf1* mutants.[35] Conceivably, some of these replisome defects could result in irrepairable DNA lesions, but how these perturbations are caused by an absence of K56 acetylation is far from clear.

HOW DOES K56 ACETYLATION PROMOTE CELL SURVIVAL IN RESPONSE TO DNA DAMAGE?

Nucleosome assembly normally takes place nearly as soon as enough DNA has been generated by the replication apparatus to allow the formation of nucleosomes.[21] Based on the fact that virtually all newly synthesized H3 molecules deposited into chromosomes during S-phase are modified, K56 acetylation is likely present close behind all replication forks.[19] This may be important to ensure that K56 acetylation is immediately accessible whenever replication forks collide with DNA lesions and irrespective of the local chromatin environment at the site of damage. However, at first glance, the ubiquitous nature of K56 acetylation during a normal S-phase seems hard to reconcile with a direct role in attracting DNA damage signalling or repair enzymes specifically to sites of stalled or damaged replisomes. Here we propose several mutually non-exclusive mechanisms by which K56 acetylation could be exploited to promote the repair of replication-linked DNA lesions.

First, unlike H2A serine 128 phosphorylation (equivalent to γH2AX in human cells), which is specifically induced at sites of DSBs and directly associates with checkpoint proteins,[61-65] K56 acetylation may not attract effector proteins to damaged replisomes by binding to them. Instead, K56 acetylation could exert its function by directly perturbing nucleosome structure. Given its location at the entry and exit points of the DNA from the nucleosome core, K56 acetylation may promote short-range histone octamer sliding along the DNA in a manner analogous to the action of Snf2-type ATP-dependent nucleosome remodeling enzymes.[66] Acetylation of K56 may also increase DNA accessibility simply by enhancing the rate of spontaneous dissociation of short DNA segments at the entry and exit point of the nucleosome.[11,12] A number of pathways can be employed to resume DNA synthesis when the replisome is blocked by DNA damage. In principle, K56 acetylation-dependent exposure of DNA segments behind stalled replication forks could facilitate mechanisms such as fork regression and/or error-free DNA lesion bypass mediated by template strand switching.[60,67] In the absence of K56 acetylation, channeling of stalled replication forks towards one of these pathways may result in unrepairable lesions.

Another possibility for a direct role of K56 acetylation is that it might disrupt the folding of chromatin into higher order structures.[68] Perhaps genome-wide K56 acetylation disrupts chromatin structure in a manner that impedes the action of condensins or other proteins required for faithful chromosome segregation. A destabilizing effect of K56 acetylation on the higher-order structure of chromatin might also help to explain why K56 acetylation is largely removed prior to mitotic chromosome segregation. Conversely, in the absence of K56 acetylation, an overly rigid chromatin higher-order structure may not be compatible with the roles of cohesins or the structurally related Smc5-Smc6 complex in DNA repair.[69-71] The *S. cerevisiae* Hho1 protein is structurally related to higher eukaryotic histone H1,[72] which is important to stabilize the higher-order structure of chromatin.[73] Interestingly, Hho1 inhibits HR-mediated DSB repair in yeast,[74] although whether this effect is mediated through folding of chromatin into a stable higher-order structure is not known.

A second type of model proposes that, while K56-acetylated H3 is deposited throughout the genome during S-phase, it may only be accessible to bind effector proteins at sites of replication fork damage. In this case, K56 acetylation would serve as a ubiquitous mark that is conditionally exposed only when needed at stalled or collapsed DNA replication forks. This is plausible because several H3 residues near the N-terminal tips of the αN helices that contain K56 make very strong contacts with the middle portion of nucleosomal DNA (Fig. 1, contacts between green residues and SHL -0.5 and +0.5). DNA segments at these points in the nucleosome are relatively inaccessible to restriction enzymes.[75,76] Therefore, the recognition of the αN helix by proteins that can exploit K56 acetylation to promote repair may require prior remodeling of these strong contacts between histone and DNA. Conceivably, several ATP-dependent nucleosome remodeling enzymes that have been implicated in the DNA damage response[77] could perform the function of exposing acetylated K56 specifically at damaged replication forks. A related model invoking regulated accessibility posits that K56 acetylation directly recruits effector proteins to damaged replication forks, but that it cannot perform this function when the acetylation is present throughout the genome during S-phase. The existence of a mechanism that only protects K56 acetylation locally at sites of damage would ensure that acetylation is removed throughout undamaged regions of the genome during S-phase and G_2, thereby eventually allowing H3 K56 acetylation to act directly in the recruitment of repair proteins to sites of damage. The latter model could explain why constitutive K56 acetylation throughout the gemone results in extremely high DNA damage sensitivity,[19] as the factors that mediate DNA repair would be 'titrated out' when K56 acetylation is present throughout the genome. A refinement of the conditional exposure model stipulates that K56 acetylation can only attract DNA repair enzymes when present in or near nucleosomes that contain a second modification which occurs exclusively at sites of replication fork damage. Although H2A serine 128 phosphorylation is an appealing candidate, the fact that H3 K56R mutants are far more sensitive than H2A S128A mutant cells to several genotoxic agents strongly argues that H3 K56 acetylation can promote DNA damage survival in a manner that is largely independent of H2A serine 128 phosphorylation.[4] However, it seems possible that future studies could uncover other histone modifications that are specifically targeted to nucleosomes near sites of replication fork damage. These modifications may well function in a combinatorial manner with H3 K56 acetylation to facilitate the recruitment of DNA repair and/or chromatin remodeling enzymes specifically to damaged forks.

CONCLUSION AND PERSPECTIVE

Although the detailed molecular mechanisms are not known, cell cycle-regulated acetylation and deacetylation of histone H3 lysine 56 both have profound impacts on the ability of cells to survive DNA lesions that halt replication fork progression. Chromosome translocations and other rearrangements are recurrent features of many human cancers.[78] Many of these chromosomal aberrations are triggered in

response to spontaneous or genotoxic agent-induced replication fork damage.[78] In addition, many clinically relevant cancer chemotherapeutic agents act by interfering with replication fork progression and causing DNA strand breaks during S-phase.[79] Cells without K56 acetylation, such as *asf1* mutants, have a high incidence of spontaneous DNA damage and chromosome rearrangements,[14,51,80,81] suggesting that a lack of K56 acetylation compromises the fidelity of DNA repair. Although there is currently no published evidence that K56 acetylation exists in human cells,[5,7] the rapid deposition of histones behind replication forks is conserved in higher eukaryotes.[21,82] Thus, the enzymes that acetylate and deacetylate histones in the vicinity of damaged replication forks may represent novel targets for cancer chemotherapy.[83]

References

1. Mersfelder EL, Parthun MR. The tale beyond the tail: Histone core domain modifications and the regulation of chromatin structure. Nucleic Acids Res 2006; 34:2653-62.
2. Zhang L, Eugeni EE, Parthun MR, Freitas MA. Identification of novel histone post-translational modifications by peptide mass fingerprinting. Chromosoma 2003; 112:77-86.
3. Hyland EM, Cosgrove MS, Molina H, Wang D, Pandey A, Cottee RJ, Boeke JD. Insights into the role of histone H3 and histone H4 core modifiable residues in *Saccharomyces cerevisiae*. Mol Cell Biol 2005; 25:10060-70.
4. Masumoto H, Hawke D, Kobayashi R, Verreault A. A role for cell-cycle-regulated histone H3 lysine 56 acetylation in the DNA damage response. Nature 2005; 436:294-8.
5. Ozdemir A, Spicuglia S, Lasonder E, Vermeulen M, Campsteijn C, Stunnenberg HG, Logie C. Characterization of lysine 56 of histone H3 as an acetylation site in *Saccharomyces cerevisiae*. J Biol Chem 2005; 280:25949-52.
6. Recht J, Tsubota T, Tanny JC, Diaz RL, Berger JM, Zhang X, Garcia BA, Shabanowitz J, Burlingame AL, Hunt DF, Kaufman PD, Allis CD. Histone chaperone Asf1 is required for histone H3 lysine 56 acetylation, a modification associated with S phase in mitosis and meiosis. Proc Natl Acad Sci USA 2006; 103:6988-93.
7. Xu F, Zhang K, Grunstein M. Acetylation in histone *H3* globular domain regulates gene expression in yeast. Cell 2005; 121:375-85.
8. Zhou H, Madden BJ, Muddiman DC, Zhang Z. Chromatin assembly factor 1 interacts with histone H3 methylated at lysine 79 in the processes of epigenetic silencing and DNA repair. Biochemistry 2006; 45:2852-61.
9. Sipiczki M. Where does fission yeast sit on the tree of life? Genome Biol 2000; 1:1011.
10. Davey CA, Sargent DF, Luger K, Maeder AW, Richmond TJ. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9Å resolution. J Mol Biol 2002; 319:1097-113.
11. Li G, Levitus M, Bustamante C, Widom J. Rapid spontaneous accessibility of nucleosomal DNA. **Nat Struct Mol Biol 2005; 12:46-53.**
12. Tomschik M, Zheng H, van Holde K, Zlatanova J, Leuba SH. **Fast, long-range, reversible** conformational fluctuations in nucleosomes revealed by single-pair fluorescence resonance energy transfer. Proc Natl Acad Sci USA 2005; 102:3278-83.
13. Adkins MW, Tyler JK. The histone chaperone Asf1p mediates global chromatin disassembly in vivo. J Biol Chem 2004; 279:52069-74.
14. Prado F, Cortes-Ledesma F, Aguilera A. The absence of the yeast chromatin assembly factor Asf1 increases genomic instability and sister chromatid exchange. EMBO Rep 2004; 5:497-502.
15. Maas NL, Miller KM, DeFazio LG, Toczyski DP. Cell cycle and checkpoint regulation of histone H3 K56 acetylation by Hst3 and Hst4. Mol Cell 2006; 23:109-19.
16. Koç A, Wheeler LJ, Mathews CK, Merrill GF. Hydroxyurea arrests DNA replication by a mechanism that preserves basal dNTP pools. J Biol Chem 2004; 279:223-30.
17. Petitjean A, Hilger F, Tatchell K. Comparison of thermosensitive alleles of the *CDC25* gene involved in the cAMP metabolism of *Saccharomyces cerevisiae*. Genetics 1990; 124:797-806.
18. Sia RA, Herald HA, Lew DJ. Cdc28 tyrosine phosphorylation and the morphogenesis checkpoint in budding yeast. Mol Biol Cell 1996; 7:1657-66.
19. Celic I, Masumoto H, Griffith WP, Meluh P, Cotter RJ, Boeke JD, Verreault A. The sirtuins Hst3 and Hst4p preserve genome integrity by controlling histone H3 lysine 56 deacetylation. Curr Biol 2006; 16:1280-9.
20. Zhu G, Spellman PT, Volpe T, Brown PO, Botstein D, Davis TN, Futcher B. Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. Nature 2000; 406:90-4.
21. Sogo JM, Stahl H, Koller T, Knippers R. Structure of replicating simian virus 40 minichromosomes. The replication fork, core histone segregation and terminal structures. J Mol Biol 1986; 189:189-204.
22. Neuwald AF, Landsman D. Gcn5-related histone N-acetyltransferases belong to a diverse superfamily that includes the yeast Spt10 protein. Trends Biochem Sci 1997; 22:154-5.
23. Hess D, Liu B, Roan NR, Sternglanz R, Winston F. Spt10-dependent transcriptional activation in *Saccharomyces cerevisiae* requires both the Spt10 acetyltransferase domain and Spt21. Mol Cell Biol 2004; 24:135-43.
24. Eriksson PR, Mendiratta G, McLaughlin NB, Wolfsberg TG, Marino-Ramirez L, Pompa TA, Jainerin M, Landsman D, Shen CH, Clark DJ. Global regulation by the yeast Spt10 protein is mediated through chromatin structure and the histone upstream activating sequence elements. Mol Cell Biol 2005; 25:9127-37.
25. Mendiratta G, Eriksson PR, Shen CH, Clark DJ. The DNA-binding domain of the yeast Spt10p activator includes a zinc finger that is homologous to foamy virus integrase. J Biol Chem 2006; 281:7040-8.
26. Parthun MR, Widom J, Gottschling DE. The major cytoplasmic histone acetyltransferase in yeast: Links to chromatin replication and histone metabolism. Cell 1996; 87:85-94.
27. Sklenar AR, Parthun MR. Characterization of yeast histone H3-specific type B histone acetyltransferases identifies an Ada2-independent Gcn5p activity. BMC Biochem 2004; 5:11.
28. Brachmann CB, Sherman JM, Devine SE, Cameron EE, Pillus L, Boeke JD. The *SIR2* gene family, conserved from bacteria to humans, functions in silencing, cell cycle progression, and chromosome stability. Genes Dev 1995; 9:2888-902.
29. Budd ME, Tong AH, Polaczek P, Peng X, Boone C, Campbell JL. A network of multi-tasking proteins at the DNA replication fork preserves genome stability. **PLoS Genet 2005; 1:** e61.
30. Pan X, Ye P, Yuan DS, Wang X, Bader JS, Boeke JD. **A DNA integrity network in the yeast** *Saccharomyces cerevisiae*. Cell 2006; 124:1069-81.
31. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C. Global mapping of the yeast genetic interaction network. Science 2004; 303:808-13.
32. Lewis LK, Storici F, Van Komen S, Calero S, Sung P, Resnick MA. Role of the nuclease activity of *Saccharomyces cerevisiae* Mre11 in repair of DNA double-strand breaks in mitotic cells. Genetics 2004; 166:1701-13.
33. Llorente B, Symington LS. The Mre11 nuclease is not required for 5' to 3' resection at multiple HO-induced double-strand breaks. Mol Cell Biol 2004; 24:9682-94.
34. Tyler JK, Adams CR, Chen SR, Kobayashi R, Kamakaka RT, Kadonaga JT. The RCAF complex mediates chromatin assembly during DNA replication and repair. Nature 1999; 402:555-60.
35. Franco AA, Lam WM, Burgers PM, Kaufman PD. Histone deposition protein Asf1 maintains DNA replisome integrity and interacts with replication factor C. Genes Dev 2005; 19:1365-75.
36. Green EM, Antczak AJ, Bailey AO, Franco AA, Wu KJ, Yates IIIrd JR, Kaufman PD. Replication-independent histone deposition by the HIR complex and Asf1. Curr Biol 2005; 15:2044-9.
37. Schermer UJ, Korber P, Hörz W. Histones are incorporated in trans during reassembly of the yeast *PHO5* promoter. Mol Cell 2005; 19:279-85.
38. Tagami H, Ray-Gallet D, Almouzni G, Nakatani Y. Histone H3.1 and H3.3 complexes mediate nucleosome assembly pathways dependent or independent of DNA synthesis. Cell 2004; 116:51-61.
39. Tyler JK, Collins KA, Prasad-Sinha J, Amiott E, Bulger M, Harte PJ, Kobayashi R, Kadonaga JT. Interaction between the *Drosophila* CAF-1 and ASF1 chromatin assembly factors. Mol Cell Biol 2001; 21:6574-84.
40. Zabaronick SR, Tyler JK. The histone chaperone anti-silencing function 1 is a global regulator of transcription independent of passage through S phase. Mol Cell Biol 2005; 25:652-60.
41. Linger J, Tyler JK. The yeast histone chaperone chromatin assembly factor 1 protects against double-strand DNA-damaging agents. Genetics 2005; 171:1513-22.
42. Emili A, Schieltz DM, Yates IIIrd JR, Hartwell LH. Dynamic interaction of DNA damage checkpoint protein Rad53 with chromatin assembly factor Asf1. Mol Cell 2001; 7:13-20.
43. Hu F, Alcasabas AA, Elledge SJ. Asf1 links Rad53 to control of chromatin assembly. Genes Dev 2001; 15:1061-6.
44. Sweeney FD, Yang F, Chi A, Shabanowitz J, Hunt DF, Durocher D. *Saccharomyces cerevisiae* Rad9 acts as a Mec1 adaptor to allow Rad53 activation. Curr Biol 2005; 15:1364-75.
45. Meijsing SH, Ehrenhofer-Murray AE. The silencing complex SAS-I links histone acetylation to the assembly of repressed chromatin by CAF-1 and Asf1 in *Saccharomyces cerevisiae*. Genes Dev 2001; 15:3169-82.
46. Osada S, Sutton A, Muster N, Brown CE, Yates IIIrd JR, Sternglanz R, Workman JL. The yeast SAS (something about silencing) protein complex contains a MYST-type putative acetyltransferase and functions with chromatin assembly factor *Asf1*. Genes Dev 2001; 15:3155-68.
47. Mousson F, Lautrette A, Thuret JY, Agez M, Courbeyrette R, Amigues B, Becker E, Neumann JM, Guerois R, Mann C, Ochsenbein F. Structural basis for the interaction of Asf1 with histone H3 and its functional implications. Proc Natl Acad Sci USA 2005; 102:5975-80.
48. Sutton A, Shia WJ, Band D, Kaufman PD, Osada S, Workman JL, Sternglanz R. Sas4 and Sas5 are required for the histone acetyltransferase activity of Sas2 in the SAS complex. J Biol Chem 2003; 278:16887-92.
49. Symington LS. Role of *RAD52* epistasis group genes in homologous recombination and double-strand break repair. Microbiol Mol Biol Rev 2002; 66:630-70.
50. Lewis LK, Karthikeyan G, Cassiano J, Resnick MA. Reduction of nucleosome assembly during new DNA synthesis impairs both major pathways of double-strand break repair. Nucleic Acids Res 2005; 33:4928-39.
51. Ramey CJ, Howar S, Adkins M, Linger J, Spicer J, Tyler JK. Activation of the DNA damage checkpoint in yeast lacking the histone chaperone anti-silencing function 1. Mol Cell Biol 2004; 24:10313-27.

52. Bird AW, Yu DY, Pray-Grant MG, Qiu Q, Harmon KE, Megee PC, Grant PA, Smith MM, Christman MF. Acetylation of histone H4 by Esa1 is required for DNA double-strand break repair. Nature 2002; 419:411-5.
53. Downs JA, Allard S, Jobin-Robitaille O, Javaheri A, Auger A, Bouchard N, Kron SJ, Jackson SP, Cote J. Binding of chromatin-modifying activities to phosphorylated histone H2A at DNA damage sites. Mol Cell 2004; 16:979-90.
54. Murr R, Loizou JI, Yang YG, Cuenin C, Li H, Wang ZQ, Herceg Z. Histone acetylation by Trrap-Tip60 modulates loading of repair proteins and repair of DNA double-strand breaks. Nat Cell Biol 2006; 8:91-9.
55. Qin S, Parthun MR. Recruitment of the type B histone acetyltransferase Hat1p to chromatin is linked to DNA double-strand breaks. Mol Cell Biol 2006; 26:3649-58.
56. Tamburini BA, Tyler JK. Localized histone acetylation and deacetylation triggered by the homologous recombination pathway of double-strand DNA repair. Mol Cell Biol 2005; 25:4903-13.
57. Qin S, Parthun MR. Histone H3 and the histone acetyltransferase Hat1p contribute to DNA double-strand break repair. Mol Cell Biol 2002; 22:8353-65.
58. Takahashi TS, Wigley DB, Walter JC. Pumps, paradoxes and ploughshares: Mechanism of the MCM2-7 DNA helicase. Trends Biochem Sci 2005; 30:437-44.
59. Byun TS, Pacek M, Yee MC, Walter JC, Cimprich KA. Functional uncoupling of MCM helicase and DNA polymerase activities activates the ATR-dependent checkpoint. Genes Dev 2005; 19:1040-52.
60. Branzei D, Foiani M. The Rad53 signal transduction pathway: Replication fork stabilization, DNA repair, and adaptation. Exp Cell Res 2006; 312:2654-9.
61. Lou Z, Minter-Dykhouse K, Franco S, Gostissa M, Rivera MA, Celeste A, Manis JP, van Deursen J, Nussenzweig A, Paull TT, Alt FW, Chen J. MDC1 maintains genomic stability by participating in the amplification of ATM-dependent DNA damage signals. Mol Cell 2006; 21:187-200.
62. Nakamura TM, Du LL, Redon C, Russell P. Histone H2A phosphorylation controls Crb2 recruitment at DNA breaks, maintains checkpoint arrest, and influences DNA repair in fission yeast. Mol Cell Biol 2004; 24:6215-30.
63. Rogakou EP, Pilch DR, Orr AH, Ivanova VS, Bonner WM. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. J Biol Chem 1998; 273:5858-68.
64. Shroff R, Arbel-Eden A, Pilch D, Ira G, Bonner WM, Petrini JH, Haber JE, Lichten M. Distribution and dynamics of chromatin modification induced by a defined DNA double-strand break. Curr Biol 2004; 14:1703-11.
65. Stucki M, Clapperton JA, Mohammad D, Yaffe MB, Smerdon SJ, Jackson SP. MDC1 directly binds phosphorylated histone H2AX to regulate cellular responses to DNA double-strand breaks. Cell 2006; 123:1213-1226.
66. Boyer LA, Logie C, Bonte E, Becker PB, Wade PA, Wolffe AP, Wu C, Imbalzano AN, Peterson CL. Functional delineation of three groups of the ATP-dependent family of chromatin remodeling enzymes. J Biol Chem 2000; 275:18864-70.
67. Osborn AJ, Elledge SJ, Zou L. Checking on the fork: The DNA replication stress-response pathway. Trends Cell Biol 2002; 12:509-516.
68. Shogren-Knaak M, Ishii H, Sun JM, Pazin MJ, Davie JR, Peterson CL. Histone H4 K16 acetylation controls chromatin structure and protein interactions. Science 2006; 311:844-7.
69. De Piccoli G, Cortes-Ledesma F, Ira G, Torres-Rosell J, Uhle S, Farmer S, Hwang JY, Machin F, Ceschia A, McAleenan A, Cordon-Preciado V, Clemente-Blanco A, Vilella-Mitjana F, Ullal P, Jarmuz A, Leitao B, Bressan D, Dotiwala F, Papusha A, Zhao X, Myung K, Haber JE, Aguilera A, Aragon L. Smc5-Smc6 mediate DNA double-strand-break repair by promoting sister-chromatid recombination. Nat Cell Biol 2006; 8:1032-1034.
70. Pebernard S, Wohlschlegel J, McDonald WH, Yates IIIrd JR, Boddy MN. The Nse5-Nse6 dimer mediates DNA repair roles of the Smc5-Smc6 complex. Mol Cell Biol 2006; 26:1617-30.
71. Strom L, Lindroos HB, Shirahige K, Sjogren C. Postreplicative recruitment of cohesin to double-strand breaks is required for DNA repair. Mol Cell 2004; 16:1003-15.
72. Ali T, Coles P, Stevens TJ, Stott K, Thomas JO. Two homologous domains of similar structure but different stability in the yeast linker histone, Hho1p. J Mol Biol 2004; 338:139-48.
73. Thoma F, Koller T, Klug A. Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. J Cell Biol 1979; 83:403-27.
74. Downs JA, Kosmidou E, Morgan A, Jackson SP. Suppression of homologous recombination by the *Saccharomyces cerevisiae* linker histone. Mol Cell 2003; 11:1685-92.
75. Anderson JD, Thastrom A, Widom J. Spontaneous access of proteins to buried nucleosomal DNA target sites occurs via a mechanism that is distinct from nucleosome translocation. Mol Cell Biol 2002; 22:7147-57.
76. Wu C, Travers A. A 'one-pot' assay for the accessibility of DNA in a nucleosome core particle. Nucleic Acids Res 2004; 32:e122.
77. Wurtele H, Verreault A. Histone post-translational modifications and the response to DNA double-strand breaks. Curr Opin Cell Biol 2006; 18:137-44.
78. Kolodner RD, Putnam CD, Myung K. Maintenance of genome stability in *Saccharomyces cerevisiae*. Science 2002; 297:552-7.
79. Hurley LH. DNA and its associated processes as targets for cancer therapy. Nat Rev Cancer 2002; 2:188-200.
80. Myung K, Pennaneach V, Kats ES, Kolodner RD. *Saccharomyces cerevisiae* chromatin assembly factors that act during DNA replication function in the maintenance of genome stability. Proc Natl Acad Sci USA 2003; 100:6640-5.
81. Schmidt KH, Wu J, Kolodner RD. Control of translocations between highly diverged genes by *Sgs1*, the *Saccharomyces cerevisiae* homolog of the Bloom's syndrome protein. Mol Cell Biol 2006; 26:5406-20.
82. Shibahara K, Stillman B. Replication-dependent marking of DNA by PCNA facilitates CAF-1-coupled inheritance of chromatin. Cell 1999; 96:575-85.
83. Fraga MF, Esteller M. Towards the human cancer epigenome. Cell Cycle 2005; 4:1377-81
84. Haase SB, Reed SI. Improved flow cytometric analysis of the budding yeast cell cycle. Cell Cycle 2002; 1:132-6.
85. English CM, Adkins CW, Carson JJ, Chruchill MEA, Tyler JK. Structural basis for the histone chaperone activity of Asf1. Cell 2006; 127:495-508.

Chapter 4

High resolution mapping of Twist to DNA in *Drosophila* embryos: Efficient functional analysis and evolutionary conservation.

Ozdemir A, Fisher-Aylor KI, Pepke S, Samanta M, Dunipace L, McCue K, Zeng L, Ogawa N, Wold BJ, and Stathopoulos A.

Research

High resolution mapping of Twist to DNA in *Drosophila* embryos: Efficient functional analysis and evolutionary conservation

Anil Ozdemir,[1,5] Katherine I. Fisher-Aylor,[1,5] Shirley Pepke,[2] Manoj Samanta,[3] Leslie Dunipace,[1] Kenneth McCue,[1] Lucy Zeng,[4] Nobuo Ogawa,[4] Barbara J. Wold,[1,6] and Angelike Stathopoulos[1,6]

[1]*Division of Biology, California Institute of Technology, Pasadena, California 91125, USA;* [2]*Center for Advanced Computing Research, California Institute of Technology, Pasadena, California 91125, USA;* [3]*Systemix Institute, Redmond, Washington 98053, USA;* [4]*Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

Cis-regulatory modules (CRMs) function by binding sequence specific transcription factors, but the relationship between in vivo physical binding and the regulatory capacity of factor-bound DNA elements remains uncertain. We investigate this relationship for the well-studied Twist factor in *Drosophila melanogaster* embryos by analyzing genome-wide factor occupancy and testing the functional significance of Twist occupied regions and motifs within regions. Twist ChIP-seq data efficiently identified previously studied Twist-dependent CRMs and robustly predicted new CRM activity in transgenesis, with newly identified Twist-occupied regions supporting diverse spatiotemporal patterns (>74% positive, $n = 31$). Some, but not all, candidate CRMs require Twist for proper expression in the embryo. The Twist motifs most favored in genome ChIP data (in vivo) differed from those most favored by Systematic Evolution of Ligands by EXponential enrichment (SELEX) (in vitro). Furthermore, the majority of ChIP-seq signals could be parsimoniously explained by a CABVTG motif located within 50 bp of the ChIP summit and, of these, CACATG was most prevalent. Mutagenesis experiments demonstrated that different Twist E-box motif types are not fully interchangeable, suggesting that the ChIP-derived consensus (CABVTG) includes sites having distinct regulatory outputs. Further analysis of position, frequency of occurrence, and sequence conservation revealed significant enrichment and conservation of CABVTG E-box motifs near Twist ChIP-seq signal summits, preferential conservation of ±150 bp surrounding Twist occupied summits, and enrichment of GA- and CA-repeat sequences near Twist occupied summits. Our results show that high resolution in vivo occupancy data can be used to drive efficient discovery and dissection of global and local *cis*-regulatory logic.

[Supplemental material is available for this article. The microarray data from this study have been submitted to the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession no. GSE26285, and the sequence data from this study have been submitted to the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under accession no. SRA027330.]

In animal genomes, *cis*-acting regulatory modules (CRMs) average ~300–500 bp in size and typically contain one or more binding motif instances for several transcription factors (Davidson 2006). DNA binding motif instances can now be readily mapped *in silico* by similarity to a consensus binding motif that has been defined through in vitro methods, or they can be derived from careful functional dissection of a few well-studied CRMs. However, many transcription factors recognize short sequence motifs that occur so frequently in the genome that virtually all gene loci have one or more, raising questions about which of these sites is occupied in the cell and what regulatory impact that occupancy has. We also know that binding motifs in the best-studied CRMs are often clustered (e.g., Ip et al. 1992a; Small et al. 1992; Berman et al. 2002; Markstein et al. 2002), presumably to facilitate coordinated and cooperative interaction among factors and cofactors and to achieve specificity relative to isolated single motif occurrences. However, we do not yet understand the logic by which motif combinations specify the functional output of the vast majority of CRMs in the genome (e.g., Lusk and Eisen 2010), and efficient identification and analysis of many more CRMs are needed to uncover these principles.

Advances in identifying candidate CRMs are coming from whole-genome approaches in which either chromatin immunoprecipitation (ChIP) is employed to find the region of DNA bound by a given transcription factor in vivo (e.g., Zeitlinger et al. 2007; Zinzen et al. 2009), or high-throughput screening assays are utilized to identify promoter and CRM functions (e.g., Landolin et al. 2010; Nam et al. 2010), although the latter have not yet been widely applied. Global ChIP assays also allow one to define de novo or refine binding motifs used by a factor in vivo and to compare this with in vitro defined motifs. ChIP-seq is a particular form of genome-wide chromatin immunoprecipitation, which can produce high positional resolution of observable DNA binding in vivo (Johnson et al. 2007). In particular, the resolution of ChIP-seq data can be used to infer, within a given binding region, which

[5]**These authors contributed equally to this work.**
[6]**Corresponding authors.**
E-mail angelike@caltech.edu.
E-mail woldb@caltech.edu.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.104018.109.

21:566–577 © 2011 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/11; www.genome.org

specific motif occurrence is likely to account for the majority of the observed ChIP signal (Valouev et al. 2008). We refer to the motif instances most likely to drive observed binding as candidate "explanatory" sites, and we explore the value of making explanatory site models for all ChIP signals to guide detailed functional assays.

We apply ChIP-seq and ChIP-chip analyses to Twist, a key transcription factor in the dorsal-ventral (DV) patterning network of the *Drosophila* early embryo. Patterning the DV axis depends partly on Twist, a bHLH transcription factor present at high levels in ventral regions of the embryo (for review, see Chopra and Levine 2009; Reeves and Stathopoulos 2009). Many previous studies have contributed to the current picture of a developmental gene network that describes embryonic DV patterning, in which more than 50 genes and 30 CRMs have been linked (for review, see Stathopoulos and Levine 2005). Previous published ChIP-chip studies conducted using Twist antibodies have demonstrated that its occupancy can be detected in vivo (Sandmann et al. 2007; Zeitlinger et al. 2007). Our goals are to relate the global Twist occupancy pattern to functional CRM activity, as assayed by transgenesis, and to relate the local ChIP-seq profile to specific motif instances and combinations and their contribution to individual CRM activity.

Results

Comparison of ChIP-chip and ChIP-seq in the identification of CRMs

We performed ChIP-chip and ChIP-seq analysis on sheared chromatin isolated from *Drosophila* embryos from 1 to 3 h in age, using an antibody that is specific to Twist protein, and subsequently assessed the overlap between sets of regions identified by each approach (see Supplemental Fig. 1A–C and Methods). For ChIP-chip, we used a script to call peaks based on a minimum signal score, whereas for ChIP-seq, we used the ERANGE software suite to call peaks based on the number, orientation, and ratio of short sequence reads relative to a background control. The results from these methods were compared at several sensitivity thresholds to accommodate different numbers of peaks called by their informatics pipelines (Supplemental Fig. 1D). Given the substantial technical and computational differences between ChIP-chip and ChIP-seq, the fact that the vast majority of ChIP-seq signals overlap with some ChIP-chip regions lends mutual confidence, although a large number of ChIP-chip sites lacked support from ChIP-seq. Inspection of multiple ChIP-seq regions for which Twist activity was previously studied in detail showed that ChIP-seq regions are generally better resolved and provide superior guidance for experimental tests of function that are the central focus of this study (Supplemental Table 1).

Functional analysis of Twist-occupied regions

We quantified how frequently and strongly ChIP-seq regions function as CRMs at the same time and place in development as the ChIP assays. To first identify a set of known gold-standard Twist CRMs, we applied a conservative standard that allowed only CRMs having prior genetic and molecular evidence. Enhancers (i.e., CRMs supporting gene expression rather than acting as silencers) along the DV axis were categorized as three types: Type I (ventral regions), Type II (ventro-lateral regions), and Type III (dorsal-lateral and dorsal regions) (Supplemental Table 2B; for review, see Chopra and Levine 2009; Reeves and Stathopoulos 2009). Many enhancers of Types I and II require Twist for expression based on genetic and molecular genetic evidence, but not until recent ChIP-chip analyses was it thought that Twist might function to regulate Type III patterns (Zeitlinger et al. 2007). We observed very strong ChIP signals at *sog* and *brk* Type III CRMs but not at *ind, dpp, zen,* and *tld* (Supplemental Table 2B; Supplemental Fig. 2). When only Type I and II CRMs were considered, 11 of 15 were present in our medium confidence (MC) data set (see Methods). Known CRMs for the four not present (i.e., *Ady43A*, *phm*, *E(spl),* and *wntD*) had below-threshold or no Twist ChIP-seq signal. The threshold for calling peaks could, of course, be reduced in order to recapture some (e.g., *wntD* and *phm*), but at the expense of increasing the false positive rate. Taken at face value, this gold standard comparison suggests we miss ~25% of true positives at the threshold selected.

Next, we tested 31 new candidate Twist CRMs drawn from the entire ChIP-seq set in a standard reporter gene assay (see Supplemental Table 2A). Of the 31 test regions, 23 (74%) supported expression; 21 supported expression in a classic dorso-ventral pattern or a subregion thereof, and 2 supported distinct patterns (i.e., ubiquitous or purely anterior-posterior) (Supplemental Fig. 3). The 23 new CRMs were distributed throughout the ChIP-seq signal range (Supplemental Fig. 2, "Positive signal"). Peaks near genes *Cyp310a1*, *Traf4, mirror (mirr)*, and *Mef2* were clearly defined by the ChIP-seq data, while the equivalent ChIP-chip data in these regions was much broader and, in some cases, gave multiple peaks, making the location of a candidate CRM ambiguous (see Fig. 1A–D). While Twist ChIP-seq data led to a high recovery rate of CRM detection, surprisingly, only ~25% of the associated genes including *Cyp310a1*, *Asph*, and *emc* (i.e., 3 of 12 assayed) actually required Twist to support expression in embryos. For instance, *mirr*, *Traf4,* and *Mef2* expression was unaffected in *twist* mutants, even though their Twist-ChIP-seq signals were equally prominent and numerous (data not shown; see Discussion).

Twist recognition motifs in vivo and in vitro

Twist belongs to a large bHLH family of DNA-binding factors that recognize a core DNA consensus, CANNTG, called an E-box (for review, see Massari and Murre 2000). Prior work using in vitro and in vivo approaches highlighted a subfamily preferred by Twist, led by CATATG (i.e., TA E-box). We asked which, if any, of the 10 possible E-box recognition motifs (counting reverse complements as the same motif) are selectively concentrated within 50 bp of called ChIP-seq signal summits (Fig. 2A). We found that CA and GA core E-boxes were most prominent, while GC, TA, and CG were relatively minor (Fig. 2A, "Twist ChIP-seq"). Compared with regions sampled from ChIP-seq control data or from the entire non-repeat genome, only CA, TA, CG, and GA core E-boxes were statistically enriched in Twist-occupied regions (Fig. 2A, colored slices). When larger radii from the ChIP signal summits were interrogated, the number of E-boxes of all types increased, and the specific enrichment trend was less apparent (i.e., enrichment of CA, TA, CG, and GA core E-boxes). In contrast, when ChIP-chip regions were similarly examined (Supplemental Figs. 5, 6), no specific enrichment of any motif was detected at any radius from the called Twist peaks. Overall, the enrichment and resolution results suggest that the ChIP-seq data could be used to model individual binding domains and causal motif instances in them (see below).

Previously published foot-printing data and small-scale SELEX had found that the in vitro Twist protein binding consensus is CAYRTG (i.e., core E-box residues YR = TA, CG, and CA) (Ip et al. 1992b; Zinzen et al. 2006). To test how Twist in vivo binding results

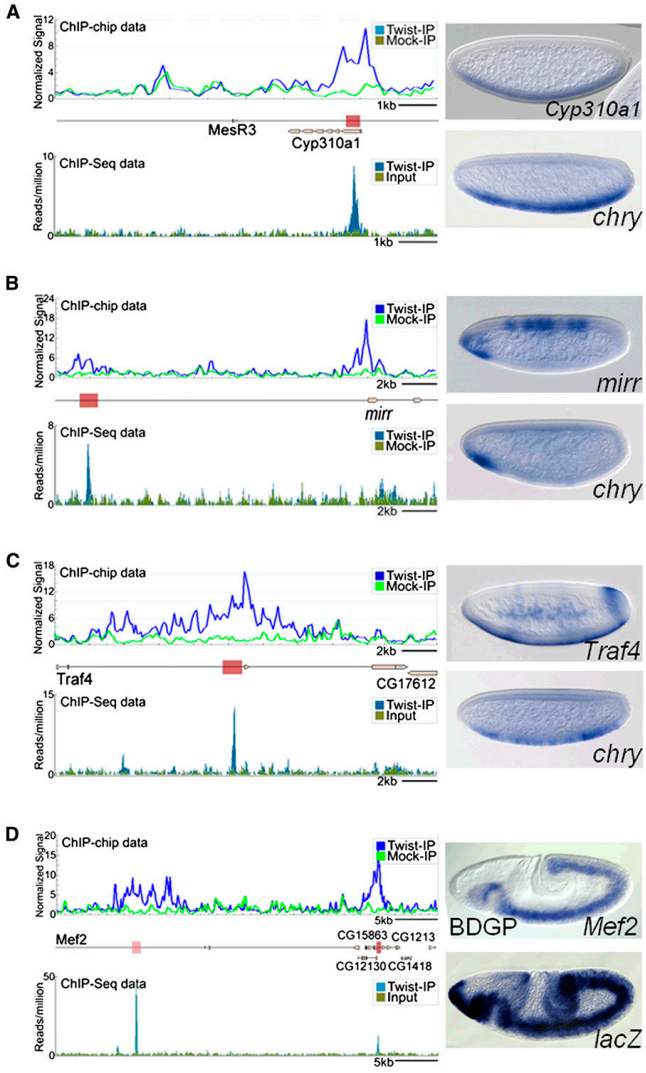


Figure 1. In vivo Twist occupancy supported by Twist ChIP-seq identifies functional CRMs. Representative examples of newly identified enhancers (brown boxes) and those previously identified (pink boxes) are shown for *Cyp310a1* (*A*), *mirr* (*B*), *Traf4* (*C*), and *Mef2* (*D*). *Upper left* panels show ChIP-chip data and *lower left* panels show ChIP-seq data for Twist-IP and control samples. In *upper right* panels, lateral views of whole mount in situ hybridizations of the endogenous genes of stage 5–8 embryos are shown. In *lower right* panels, lateral views of whole mount in situ hybridizations of similar staged embryos containing either *cherry* (for *Traf4*, *mirr*, and *Cyp310a1* enhancers) or *lacZ* (for *Mef2 5′* enhancer) reporter constructs.

relate to in vitro preferences, we determined E-box frequencies in high-throughput Twist SELEX data, and compared them with our ChIP-seq data (see Supplemental Text). For the most part, the same E-boxes were highlighted, except that the TA-core E-box motif, which was the most highly bound by Twist in vitro (35.6% occupancy by SELEX), was less enriched in vivo (7% by ChIP-seq versus 5.3% frequency in the genome). A simple explanation is that there are real differences between the in vivo and in vitro binding conditions that affect Twist motif preference. Among alternative explanations, one or more species of bHLH heterodimers might be acting in vivo, while only homodimers were assayed in vitro (see Discussion).

Motif composition of Twist ChIP-seq regions

We examined the positions of all E-box motifs within Twist-ChIP-seq regions (Fig. 2B). The ChIP-seq protocol used here is a standard Illumina platform one that retains information about whether a sequenced fragment end originated from the Watson (red) or Crick (blue) strand (Fig. 2B; Valouev et al. 2008). With appropriate data preprocessing to account for fragment length (for review, see Pepke et al. 2009, see Methods), the summit location within each peak region can be identified computationally. Inspection of known Twist CRMs showed that this agrees well with, on average, 1–2 dominant binding motif instances within ±50 bp (e.g., see Fig. 2B). A subset of previously known Twist-bound regions consists of multiple peaks aggregated together, and these are typically associated with multiple Twist motifs (e.g., see Fig. 2B, *vnd*).

We mapped and visualized the position of each motif instance relative to its peak summit and calculated the cumulative frequency for each motif type as a function of distance from the peak (Fig. 3). Within the top ranked ~1000 peaks the concentration of CAYRTG motifs was stronger than in lower ranked peaks, with CACATG sites, rather than CACGTG and CATATG, being most prominent near peak summits (Fig. 3B, top). Several criteria, including manual inspection of peaks throughout the ranking and the presence of previously studied Twist-dependent CRMs, led us to define a high confidence (HC) threshold of 513 regions (FDR 1%; see Methods and Supplemental

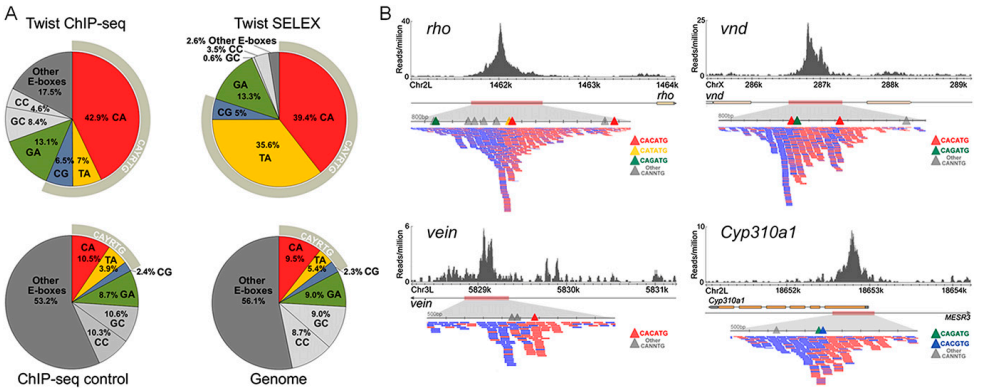


Figure 2. A comparison of Twist in vivo and in vitro binding preferences. (*A*) The frequency of E-boxes associated with HC twist peaks (±50 bp), SELEX-bound sequences, ChIP-seq enriched control regions (±50 bp of summits), and the non-repeat dm3 genome was calculated. (*B*) Twist ChIP-seq data in the vicinity of CRMs shown to support expression of the genes *rho* (Ip et al. 1992b), *vnd* (Stathopoulos et al. 2002), *vein* (Markstein et al. 2004), and *Cyp310a1* (this work). The directionality within ChIP-seq sequencing reads points to the position of the "explanatory" site. Blue and red ticks symbolize individual sequencing reads acquired, which match either the Watson or Crick strand.

Text); however we also found that binding motif centrality extends to ~1000 sites in the genome, and for most analyses we use this more inclusive set of ~1000 medium confidence (MC) calls (FDR 17%).

The accumulation of motif instances as a function of distance from the summit, over the entire set of Twist ChIP-seq regions, was analyzed (Fig. 3B, bottom). Using the K-S test, the *P*-value for CACATG distribution was defined as $<2.2 \times 10^{-16}$ (D = +0.44), meaning that the observed enrichment of CACATG near the peak summit is non-random and highly significant. It suggests that the CA-containing E-box drives in vivo binding at the majority of sites we called. Five other E-boxes also are enriched near summits, though they are less frequent in comparison to CACATG (Fig. 3B, top; Supplemental Fig. 8; Supplemental Table 3). In addition, the highest ranking peaks are associated with 2 or more matches to E-boxes; in particular the CACATG site is prominent (see Supplemental Fig.9).

CACATG and CATATG motifs are not functionally synonymous

For many ChIP regions, detailed inspection of the primary data displayed in browser format confirms a single explanatory motif (e.g., *vein* CRM, Fig. 2B; Supplemental Fig. 10). However, some CRMs contain two or more closely spaced sites matching the CABVTG consensus, leading us to ask how closely positioned E-boxes interact. The *rho* early embryonic enhancer is such a case, with a highly directional single peak with two E-boxes sites (CATATG, T1, and CACATG, T2) separated by only 5 bp (Fig. 4A). We tested whether a series of enhancer constructs support expression in the lateral domain of the embryo, comparing the wild type CRM with Twist motif mutants.

Within the *rho* enhancer sequence, we introduced single-nucleotide changes to sites T1 and T2 (CANNTG→GANNTG). These subtle changes abrogated expression, such that instead of supporting expression in a wide domain (~6–8 cells), the mutant enhancer supports expression in a more narrow domain (~3–4 cells) (cf. Fig. 4D,C); this result is comparable to what others have found previously with more severe changes to the T1 and T2 E-box sequence (5 or more changes per site; Ip et al. 1992c). We also found that mutation of either site alone supported reporter gene expression, but neither was as severe as eliminating both (cf. Fig. 4E,F,G and 4C,D). This suggested that Twist binding to both T1 and T2 sites contributes to *rho* expression.

We then asked whether CA and TA E-boxes are interchangeable. When T1 and T2 are both CACATG (i.e., T1 site TA-core was converted into CA-core), reporter expression was comparable to wild type (Fig. 4I). In contrast, replacement of both sites by the CATATG was not sufficient to support expression over the full spatial domain (Fig. 4H); in fact, expression was comparable to the T2 mutant (Fig. 4G). This suggests that the CA E-box can function in both positions, while the TA E-box can function in T1 but not T2.

Motif discovery in Twist ChIP-seq regions

To uncover possible alternative Twist binding motifs or co-associated motifs for Twist-interacting factors, we used MEME, a motif discovery tool (Bailey et al. 2006), to search for statistically over-represented motifs in and near Twist-occupied regions. As expected, prominent motifs found by MEME were E-box sequences (Fig. 5A) that overlap with CABVTG defined by our previous analyses (Fig. 3). In addition, MEME output highlighted residues flanking the E-box, such that a leading-A or lagging-T residue is preferred [e.g., CACATG-T (A-CATGTG) or A-CACATG (CATGTG-T)]. In contrast, a lagging A was very rare in Twist regions and in the genome at large (Fig. 5A). Other in vitro and in vivo bHLH binding studies support the idea that flanking bases may influence bHLH DNA binding (Grove et al. 2009; Cao et al. 2010).

Several "simple" repeat sequences were significantly over-represented in the Twist-occupied regions: the predominant one was a CA-repeat, and a similar GA-repeat sequence was also found (Fig. 5A). Of the 1099 peaks comprising the MC Twist ChIP-seq data set, 850 contain at least one match to either major E-box in the wide area around the peak (±250 bp), and 378 of these (or 44%)

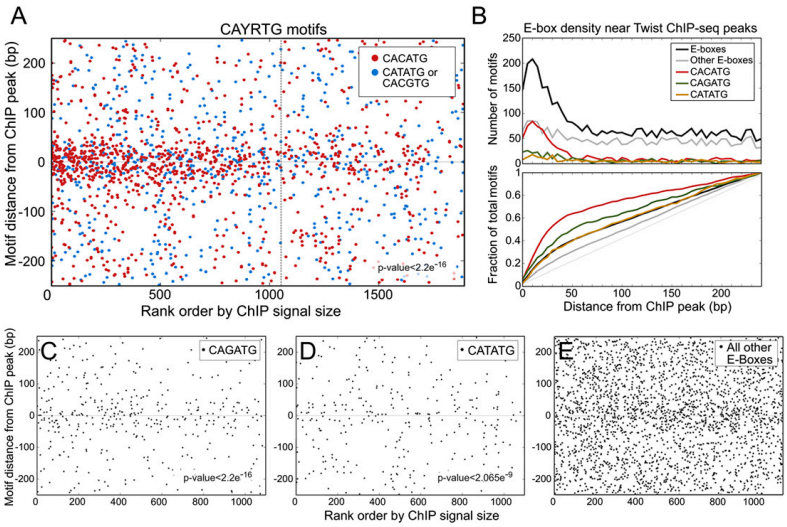


Figure 3. Motif composition of Twist ChIP-seq regions shows preferential concentration of specific E-boxes near summits. (*A*) Locations of CAYRTG = CACATG CATATG and CACGTG E-box instances located within ±250 bp of the ChIP-seq peak (ERANGE-shifted called signal summit; see Methods) (*y* axis), plotted as a function of signal intensity rank from highest (1) to lowest (2000) (*x* axis). 1099 MC ChIP-seq data set is shown with a dashed line. CACATG is the most prevalent E-box motif in Twist ChIP regions and it shows the strongest central concentration. (*B*) Direct (*top* panel) and cumulative (*bottom* panel) motif density plots. In the MC data set, 65% of CACATG motifs and 50% of CAGATG occur within ±50 bp of Twist peaks. (*C*) CAGATG occurs more frequently in Twist ChIP-seq regions and is more centrally localized than (*D*). (*D*) CATATG is the motif most prominent in SELEX data (see text). (*E*) Other E-boxes (defined here as CANNTG motifs where NN is neither CA, GA, nor TA) display a more uniform distribution (*B*,*E*), though the other CABVTG E-boxes not pictured here (CG, GC, and CC) provide a minor central enrichment (see Supplemental Fig. 8). The number and distribution of explanatory E-boxes changes with ChIP-seq signal strength, suggesting that more E-boxes create a more robust Twist ChIP signal (*A*; Supplemental Fig. 7).

also contain at least one CA- or GA-repeat sequence (Fig. 5B). It is possible that the CA- and GA-repeats associated with Twist ChIP-seq peaks play some role in marking or phasing these regions as potentially "open chromatin", as these same motifs were recently found associated with DNA occupied by Trithorax and Polycomb group/recruitment factors (see Schuettengruber and Cavalli 2009; and Discussion).

Interactions between Twist and other transcription factors might exist, yet not be identified by MEME for various reasons. We therefore tested additional motifs already known to bind transcription factors that pattern the DV axis in the early *Drosophila* embryo. Dorsal is a maternal transcription factor that functions cooperatively with Twist at some well-studied, closely-spaced sites (e.g., Ip et al. 1992c; Erives and Levine 2004), but the generality of this pattern across other Twist bound regions is not known. We found no significant global correlation between Dorsal motif occurrences and Twist peaks in our data. Among other factors (i.e., Su(H), Zelda, RGGNCAG/unknown, and Snail), only Snail exhibited significant motif co-enrichment in Twist ChIP regions, while Su(H) and RGGNCAG exhibited weak enrichment. The Snail result is neither surprising nor definitive because this factor can bind a sequence similar to that of Twist (Supplemental Fig. 12). Snail is thought to function as a repressor, at least in part, by competitively inhibiting binding of Twist (e.g., Ip et al. 1992b). Perhaps binding of both Twist and Snail to CRMs through the CA-core E-box plays a role that is more widespread than previously appreciated (see Discussion).

Twist-occupied regions were preferentially and significantly concentrated in proximal promoters (Fig. 6A), relative to several control samples, while intronic and intergenic classes were not enriched. Twist regions were slightly, but not significantly, depleted in exons. We tested whether the Twist regions near promoters were, more frequently than any others, lacking an explanatory E-box. This would be expected if many Twist promoter ChIP signals resulted from capture of indirect looping interactions from distant Twist-bound CRMS (e.g., Fullwood and Ruan 2009), rather than from primary motif binding, but it was not observed (Fig. 6B). We also asked if specific E-box motifs are selectively associated with any specific gene region class. Explanatory motifs at promoters showed higher CAGCTG and CACGTG E-box content, relative to intronic and intergenic groups, and a reduction in the dominant CACATG motif (Fig. 6B; Supplemental Fig. 13). These trends were not due to similar changes in the frequencies of GC, CG, or CA dinucleotides in promoters genome-wide (Supplemental Fig. 13). Exons also had distinctive signatures, presumably due to protein coding constraints.

Evolutionary conservation of ChIP-seq regions and motifs

Preferential sequence conservation is a signature of many biologically-significant regulatory regions and sequence motif instances. On average, our Twist-occupied regions were more conserved over a sequence domain of ~300 bp compared to random genomic background conservation (blue versus red trace, Fig. 7A).

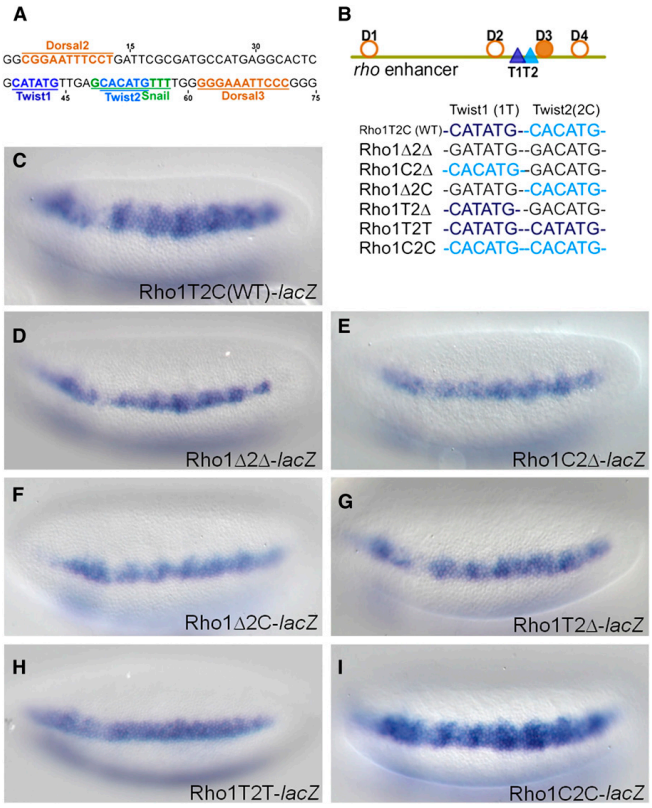


Figure 4. Mutagenesis of Twist binding sites at the ChIP-seq peak summit of *rho* enhancer. (*A*) The 75 bp sequence from the *rho* minimal enhancer which contains binding sites for Twist as well as for the transcription factors Dorsal and Snail. E-box sequences CATATG (T1, dark blue) and CACATG (T2, light blue) are separated by 5 bp, and Dorsal binding sites (orange) are positioned upstream and downstream of Twist sites. A Snail site that overlaps with T2 E-box is shown in green. (*B*) A diagram of the minimal 299 bp *rho* enhancer showing the relative positions of sites for Twist (dark and light blue triangles) and Dorsal (orange circles and filled circles, showing non-canonical and canonical sites, respectively). Lower schematic shows color-coded representations of the WT or mutant Twist binding sites present in various reporter constructs. Single nucleotide mutations were introduced into either T1 or T2 to eliminate binding (black: CATATG>GATATG or CACATG>GACATG) or to convert one site to the other (light blue: CATATG>CACATG or dark blue: CACATG>CATATG). (*C*) In situ staining of the wild type construct, minimal *rho* enhancer attached to the *evep.lacZ* reporter. (*D*) The Rho1Δ2Δ double mutant containing point mutations in both of the E-boxes, T1 and T2, supports reporter gene expression that is significantly weakened and more narrow compared to wild type (*C*). (*E–G*) Single mutations support expression that is weaker than wild type (*C*), more similar to the double mutant (*D*). (*H*) When a CATATG E-box is present in both the T1 and T2 positions, this change severely affects the expression domain of the reporter gene, reducing it to levels comparable to those observed in the double mutant Rho1Δ2Δ embryos (*D*). (*I*) When a CACATG E-box is present in both the T1 and T2 positions, the expression supported is comparable to the wild type (*C*).

In the HC Twist ChIP-seq data set of 513 peaks, conservation was highest over the motif when regions were centered on the explanatory CABVTG instance, and conservation gradually dropped to background levels as a function of distance from the center (green versus blue trace, Fig. 7A). Slight preferential conservation is observed in the background control sequence when they are aligned using the same set of E-boxes (cyan versus red trace, Fig. 7A). This is consistent with E-boxes being targets of a large family of transcription factors that exhibit varying degrees of motif preference. Furthermore, this regional conservation was less prominent in lower ranked peaks, suggesting that the higher ranked peaks are more likely to be functional (see Supplemental Fig. 14).

To assess conservation of E-box sites more quantitatively, we compared the distribution of phastCons scores for inferred Twist binding motifs in peak domains (±150 bp from the ChIP-seq summit) to those for other sequences in the same regions (Fig. 7B). E-box motifs were significantly more conserved than the rest of the domain, suggesting that they are more functionally relevant than the average sequence around them. This supports the view that E-boxes in proximity to detected peaks are not only "explanatory" for binding, but that many of these have some function in vivo. The function implied by conservation may or may not occur during the embryonic stage at which we have made our measurements, and it is even possible that some are conserved due to binding by a different bHLH factor during the life of the animal.

We examined the degree of conservation of individual E-boxes of interest relative to one another and to CA and GA repeats that were found to be prevalent in the ChIP-seq signals. We sought to distinguish those with functions associated specifically with the Twist-occupied CRMs by comparison to flanking sequence, by comparing the fraction of conserved (phastCons > 0.9) motif occurrences within ±150 bp of the ChIP-seq summit to those in flanking regions 250–500 bp away from the summit (Fig. 7C); the latter is assumed to be statistically equivalent to genomic background from data in Figure 6A. We find that CATATG, CACATG, and GA repeats stand out in terms of the change in conservation between peak and flanking sequences. In contrast, CAGATG, CACGTG, CACCTG, and CA repeats show minimal change between peak and non-peak sequences.

Discussion

This analysis of in vivo Twist occupancy in the developing *Drosophila* embryo provides general and specific insights into relationships of Twist DNA binding motifs and in vivo Twist occupancy with regulatory function. We found that the in vivo

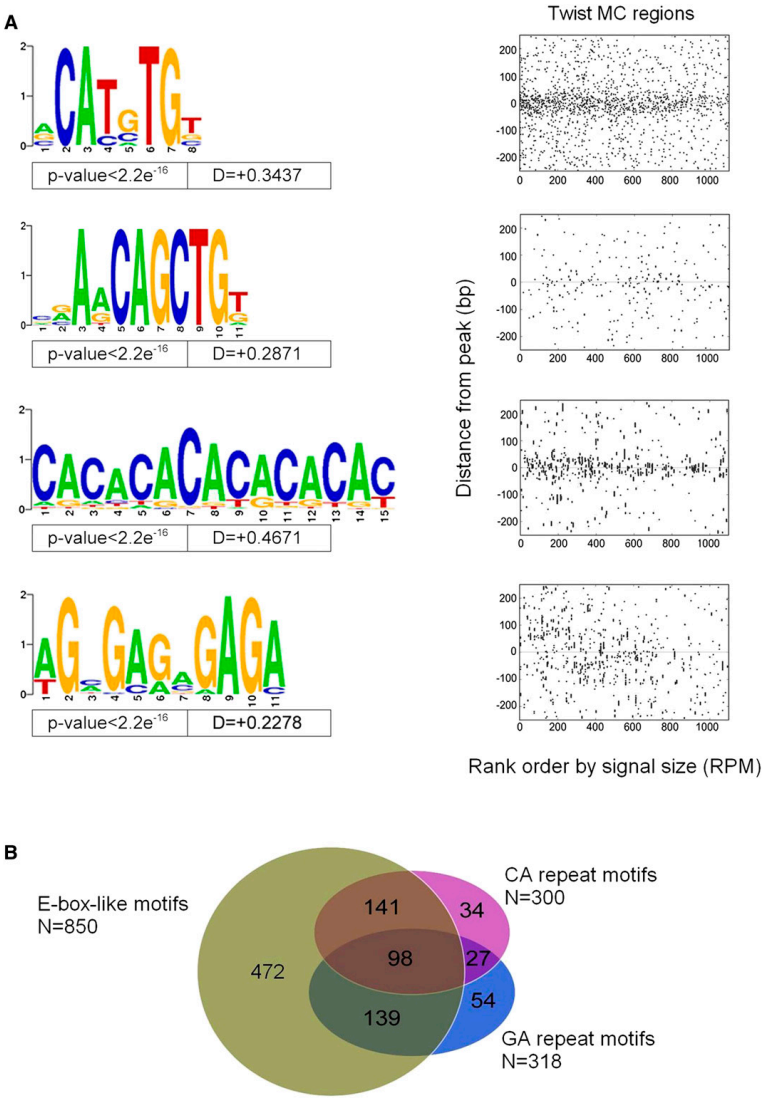


Figure 5. Motifs associated with Twist in vivo occupancy identified using MEME. MEME was run on the narrow 50 bp region surrounding each of the 1099 MC ChIP-seq peaks to identify all motifs that are enriched near the point of Twist occupancy. These motifs were mapped back to determine their spatial distribution relative to Twist peaks, and some motifs showing a non-uniform distribution near Twist peaks were selected. (*A*) Variations on CAYRTG and CAGCTG were returned, together specifying CABVTG (*top* two Weblogos). Note that a leading A residue or a lagging T residue is also suggested, which appears preferred by other non-Twist family DNA-binding bHLH factors (K Fisher-Aylor, S Kuntz, and A Kirilusha, unpubl. obs.; Grove et al. 2009). In addition, two simple repetitive sequences (CA and GA) are also significantly enriched at Twist-occupied sites (*bottom* two Weblogos). (*B*) Venn diagram illustrating the relationship between sets of peaks defined as having at least one occurrence of (i) either of the two E-box-like motifs; (ii) the CA-repeat-like sequence; or (iii) the GA-repeat-like sequence.

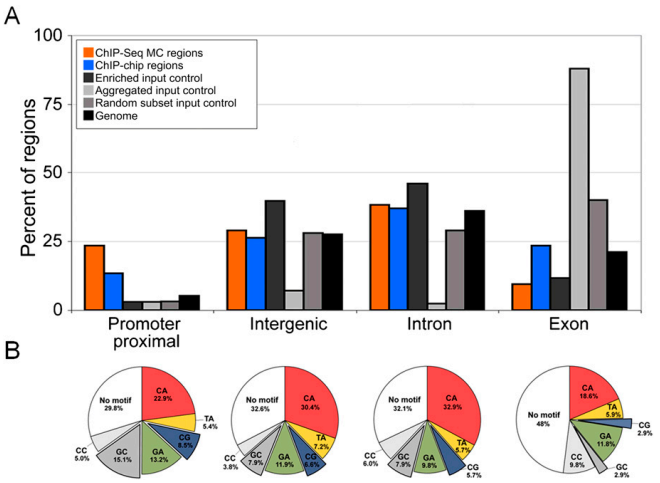


Figure 6. Enrichment of Twist ChIP-seq summits and explanatory E-box motifs in different genic and intergenic locations. (*A*) Enrichment of Twist ChIP-seq and ChIP-chip summits at particular positions in the genome, relative to a genome random sample and several sequencing negative controls. The genome was segregated into four mutually exclusive categories: promoter proximal (relative to the set of promoters from S. Celniker, including 500 bp upstream), exonic, intronic, and intergenic (see Supplemental Methods). While the majority of Twist regions fall into intergenic and intronic regions, there is a significant overabundance of Twist peaks in promoters relative to the amount of promoters in the genome (24%, or 258 of the ChIP-seq peaks). Intergenic and intronic Twist occurrences are comparable to that expected from a random genomic sample (29%, or 319 intergenic, and 38%, or 420 intronic). The number of summits within exonic regions is relatively disenriched (9%, or 102). In order to assess these numbers compared to expected values, we also compared the same number of Twist ChIP-chip regions (largest by area), the input control DNA regions enriched over Twist, the aggregated input DNA, and a random sampling of sequenced reads mapping uniquely to the genome (see Supplemental Text). We also report the total amount of the genome falling into each of these categories. The aggregated control and, to a lesser degree, the random control reads draw attention to the fact that there are many sequenced reads falling into exons. The enriched control does not show the exon bias perhaps because a directionality requirement was used; there is a mild enrichment of these sequences in the gene flanking category relative to the random genomic sample but a significant depletion in the promoter proximal that is likely due to the fact that Twist peaks are enriched at promoters. (*B*) The frequency of explanatory E-box sequences as a function of position of Twist-bound peaks in the genome (i.e., promoter proximal, intergenic, intronic, and exonic position). The CA, CG, and GA core E-boxes show enrichment in promoter, intergenic, and intronic positions; the GC core E-box is specifically enriched in the promoter proximal position.

consensus binding motif, as derived from Twist ChIP-seq data, is CABVTG (Figs. 2 and 5). Within that subfamily of E-boxes, CACATG is most prevalent within tested CRMs and is occupied preferentially within ChIP-seq defined peaks in general (Supplemental Tables 1 and 2; Fig. 3). Our detailed analysis of the *rho* enhancer showed that within the Twist-subfamily of E-boxes, individual members are not always interchangeable, and this suggests that they can support different functions (Fig. 4). When we searched for other motifs in addition to the E-box sequence that are associated with Twist peaks, we found that two repeat sequences, in particular, are associated with Twist ChIP-seq peaks, CA- and GA- repeat sequences, and that A/T-rich sequences are generally depleted from the region around ChIP signals (Supplemental Fig. 11). E-boxes and the over-represented motifs, in particular CACATG, CATATG, and a GA-repeat, are more conserved within peaks than background, suggesting that they have significant functions, presumably in transcriptional regulation.

We investigated the relationship between Twist occupancy and CRM regulatory activity by conducting functional tests and through analyses of conservation. Because the numbers of Twist-occupied sites we detected (500–1100) is large compared to the number of known Twist-regulated genes, it was not a foregone conclusion that most occupied regions would have any regulatory function. Our observed 74% CRM activity rate (23 positive CRMs of 31 tested) is high, and it argues that ChIP occupancy is efficiently highlighting functional regulatory DNA segments (Supplemental Table 2A); this analysis also captured the majority of gold standard enhancers identified by a number of previous studies (Supplemental Table 2B). Results showing preferential conservation of the Twist-bound cohort provide additional support for the idea that many other candidate regions that we did not test directly for function will also turn out to be CRMs.

A natural question is why the remaining ~25% did not score as active enhancers to support gene expression. Simple biological possibilities are that some Twist occupancy is not associated with any regulatory activity; that the module's regulatory activity is to silence or to insulate, rather than to enhance; that the module is bound but is not active at this time in development (for review, see Levine and Tjian 2003; Arnosti and Kulkarni 2005; Gurudatta and Corces 2009; Cao et al. 2010). There are precedents for all these possibilities, although not all have been explicitly shown for Twist. Technical explanations are that CRM activity might not have been successfully captured in a segment tested, or that the original ChIP region calls include an unrecognized class of false positives.

Although our ChIP data efficiently identified CRMs, we emphasize that there is a distinction between significant in vivo Twist occupancy, as indicated by the ChIP-seq data, versus significant regulatory dependence on Twist, which appears to be rarer. Lower levels of regulatory dependency are, at present, difficult to measure, and they might be common. At the extreme, Twist-binding at most CRMs could be entirely opportunistic, arising by protein-protein interactions with other already bound factors and cofactors and/or binding to an E-box that has been made accessible by other unrelated factors nearby.

Incongruity between in vivo and in vitro preferred motifs

Our findings suggest that the TA-core and CA-core E-boxes are similarly preferential for Twist binding in vitro, but in vivo the Twist ChIP-seq explanatory sites are enriched in CA-core E-boxes. If Twist protein sees CA and TA motifs similarly, then the in vivo preference might simply reflect general base composition. When we specifically tested for this, the magnitude of CA enrichment in Twist bound E-boxes was much larger than in the non-coding

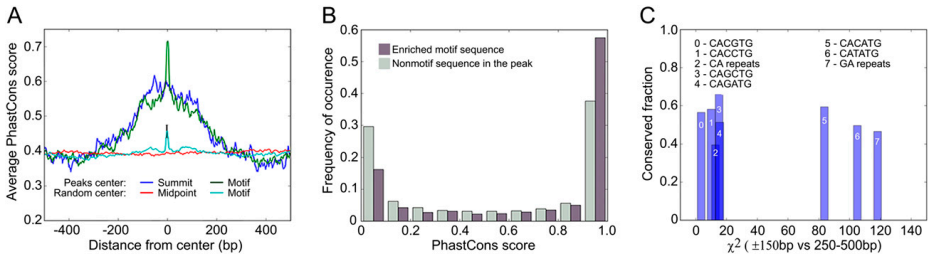


Figure 7. Conservation analysis of sequences defined by Twist binding. (*A*) Averaged conservation profiles using phastCons scores for ChIP-seq regions and random genome samples. The blue curve shows average conservation in ChIP-seq peak regions is significantly elevated ±150–200 bp from the ChIP-seq signal summit. The green curve shows the same data but with regions recentered over the nearest CABVTG binding motif within 150 bp of the original summit. For the random sample, 500 regions containing one of the motifs were selected with the region start point selected at random for the uncentered distribution. Here "midpoint" refers to the location in the center of the randomly determined region. The error bar shows two standard deviations of 30 trials of 500 samples each. A maximum over the motifs is manifest, though substantially smaller than within the ChIP-seq peak regions. (*B*) Histogram of phastCons scores for bp occurring within the 6 E-box binding motif candidates (gray) compared to that for bp within the ChIP-seq regions, but outside any of the E-box motifs (black). Bp in the motif sites are found to be statistically more conserved than bp outside of motifs (0.005 significance level). (*C*) Fraction of sites in various sequence patterns falling within the top decile of phastCons scores for a 150 bp radius surrounding ChIP-seq summits versus the chi squared statistic for distributions within 150 bp of the summit compared to those of region 250–500 bp from the summit. CACATG, CATATG, and GA repeat sequences exhibit significantly greater conservation in ChIP-seq regions compared to flanking sequence than other motifs (as shown by their clustering at high values of the chi squared statistic), though CATATG and GA repeats do not exhibit high absolute levels of conservation.

genome at large (Supplemental Fig. 13). Alternatively, bHLH proteins are known to form heterodimers in addition to homodimers, and an explanation for CA differences is that Twist binding detected in vivo is a combination of homo- and heterodimers (e.g., Murre et al. 1989). The enrichment of CA core E-boxes in vivo could reflect a particular Twist–bHLH heterodimer, since ChIP will, in principle, recover any Twist-containing complex. In particular, there is some genetic interaction data that suggests that Twist and Daughterless (Da), a bHLH ubiquitously expressed in the embryo, may interact to affect patterning in the early embryo (Jiang et al. 1992; Gonzalez-Crespo and Levine 1993; Stathopoulos and Levine 2002). Other data with forced heterodimers showed that Twist can partner with Da at later stages to influence somatic mesoderm specification (Castanon et al. 2001). When we examined overlap between our Twist ChIP-seq binding events and that of Da ChIP-chip data available (Li et al. 2008), using relaxed criteria for overlap, we found 30% of our high confidence sites have some evidence for Da binding at the same locus. When the explanatory E-box instances for these regions from our data were interrogated, we found no positive correlation with CA core E-boxes and Da, but we did find a positive correlation with GC core E-boxes and possible Twist/Da co-occupancy (data not shown). Since other bHLH factors in the embryo might also partner with Twist, the specific role, if any, of heterodimers in this system will be speculative until the full partnering repertoire for Twist is quantified and characterized. It is also possible that post-translational modifications and local conditions in the nucleus that differ from the in vitro conditions affect DNA binding preferences.

Our mutagenesis experiments with the *rho* CRM further demonstrate that the TA-core and CA-core E-boxes are not equivalent, at least in some instances. What could be different about CA- versus TA-core E-boxes? CACATG and CATATG E-boxes (e.g., T1 and T2; see Fig. 4) were first identified as Twist-binding sites within the *rho* early embryonic enhancer in 1991 by Ip et al. (1992c) using in vitro footprinting. They showed that the CA-core E-box (but not TA-core) can also be bound by the repressor Snail. It is therefore possible that the preference we see for CA core E-boxes near ChIP-seq peaks indicates that Twist/Snail combined sites have been favorably selected, and that this combination site has a distinct role in regulating the activity of many CRMs in the early embryo. In 2002, the CA-core E-box was also found to be overrepresented in a small group of CRMs that specifically support expression in ventro-lateral domains of the embryo (Stathopoulos et al. 2002), and since then others have studied cooperativity between Twist and Dorsal binding (e.g., Erives and Levine 2004; Zinzen et al. 2006; Crocker et al. 2008). It might follow that the CA-core E-box is generally required to support cooperative interactions with Dorsal or with other collaborating factors, although we did not detect Dorsal motifs in most Twist ChIP-seq defined regions.

We favor the view that in the majority of regions the Twist motif highlighted by ChIP-seq is the one most likely to contribute to regulating gene expression (or other unidentified functions), but we cannot dismiss contributions from other E-box sites present in the region. Our experiments with the *rho* enhancer illustrate this, as both E-boxes CACATG and CATATG, located five nucleotides apart, affect gene expression. Within Twist ChIP-seq peaks, we find that TA core E-boxes are less frequent overall and only weakly enriched under peaks of binding (±250 bp from the peak summit), and as a result they are not often "explanatory" (<±50 bp from the peak summit). Yet these accessory TA core E-boxes may also contribute to regulating gene expression, whether by binding Twist more transiently or by interacting with some other factor. Because the CA core E-box is also bound by Snail, the balance of activation/repression may require that a combination of CA and TA core E-boxes is optimal to support expression. Furthermore, while Twist bound to the explanatory sites may serve a major role in regulating gene expression and these accessory sites may provide less input, even marginal input may be crucial to support gene expression patterns in ways that matter for viability and selection, even though some of these may also be too subtle for our assays to detect.

Simple sequence motifs and chromatin status

Apart from the CA- and GA-repeat sequences, no motifs other than the E-boxes were found to co-cluster with Twist binding sites in

a large fraction of Twist-bound regions even when a wider window around the peaks of detected binding was interrogated. This does not preclude that other factors function in important combinations with Twist, but it suggests that no single transcription factor motif is commonly used in the entire Twist-occupied set. Finding specific combinations will require focus on subsets of regions selected by other criteria, such as expression pattern of nearby genes, performance of CRMs in transgenic assays, or direct binding assays for known or suspected accessory factors.

We do not know the significance of CA- and GA-simple repeat motifs that are enriched in Twist binding regions, but their association in other studies with open chromatin regions is suggestive (Auerbach et al. 2009). We hypothesized that GAGA-binding factor (GAF) which binds to promoters (for review, see Lehmann 2004) might do so here in promoter proximal regions through recognition of the GA-repeats. However, we did not find an enrichment of GA-repeat sequences associated with promoter proximal Twist peaks; the GA-repeats were located in many different positions suggesting a broader role than regulation of promoters, such as making DNA regions accessible.

Depletion of A/T-rich sequences from peaks was striking and it proved to be non-specific, as it is associated with a multitude of ChIP-seq samples. Further analyses showed there is a similar depletion of A/T-rich sequences around ChIP-seq peaks for diverse factors and in multiple genomes, including worm, mouse, and human (Supplemental Fig. 15; K Fisher-Aylor and B Wold, unpubl. obs.). This depletion was also seen when "peaks" of reads were selected from matching control samples of input chromatin (cross-linked, sheared, and reverse cross-linked). The sonication step associated with ChIP-seq has recently been shown to enrich for promoter regions, DNase I hypersensitive sites, and other "open" chromatin regions (Auerbach et al. 2009), but in that work no specific sequence content biases were reported. The depletion of A/T rich runs might arise from a role these sequences have been suggested to play in nucleosome exclusion and positioning (e.g., Iyer and Struhl 1995; Peckham et al. 2007). Our observations of broad A/T depletion arose from a study of motif representation that happened to be A-rich (Supplemental Fig. 11), and it suggests that careful examination of background input chromatin is needed when evaluating the sequence composition of ChIP regions.

The conservation profile around explanatory Twist motifs implies CRMs of ~300 bp

The genomes of *Drosophilids* are known to exhibit more conservation, in general, than many other animal species separated by what are thought to be an equivalent length of evolutionary distance. Thus, it has proven difficult to identify putative CRMs based on a simple search for increased local conservation of non-coding DNA sequence among *Drosophilid* genomes. Early comparative studies of enhancer regions in *Drosophila* species suggested that local increases in conservation of non-coding sequence imply regulatory function (Bergman et al. 2002). More recently, it has been suggested that this idea should be narrowed to conservation of specific binding sites only within CRMs or even just conservation of site number without strong primary sequence conservation (Sosinsky et al. 2007; Ho et al. 2009; Liberman and Stathopoulos 2009). Here we provide evidence to support both views: increased general conservation of sequence within putative CRMs relative to genomic background, as well as higher conservation of particular binding sites (Fig. 7). We asked if there is a genome-wide average conservation signature that would characterize candidate CRMs; ChIP-chip data previously detected a conservation preference but without clarity about the dimensions of regions under selective pressure (MacArthur et al. 2009). Our data suggests that sequences around these motif instances are preferentially conserved compared with genomic background in a window of ~300 bp on average, a size that corresponds well with anecdotal samplings of individual CRMs. We also found evidence that the explanatory sites identified by Twist binding are preferentially conserved compared with their surroundings, arguing for their biological salience.

Methods

Fly stocks and general molecular biology

Drosophila melanogaster fly stocks were reared under standard conditions at 25°C. Transgenic flies were obtained using standard P-element transformation or by site-directed integration. Wild type refers to the background *yw*. P-element transformations were achieved in *yw* flies, while site-directed integration was carried out using *D. mel* stock containing attP insertion at position ZH-86Fb. Enhancer sequences were amplified from genomic DNA (primer sequences are available upon request) and cloned into eve.promoter-LacZ-attB or eve.promoter-cherry-attB vectors (Liberman and Stathopoulos 2009). Anti-sense riboprobes labeled with Digoxigenin-UTP (Roche) were used for in situ hybridization to detect transcripts.

Chromatin preparation, DNA isolation, amplification, hybridization, and sequencing

Chromatin was prepared as described previously (Sandmann et al. 2006) from 2 g of *yw* embryos of from 1 to 3 h in age. Rat anti-Twist antibody (gift of M. Levine, UC Berkeley) was used for both ChIP-chip and ChIP-seq experiments. For ChIP-chip, the resulting DNA library was labeled and hybridized to arrays by NimbleGen Systems, Inc.; 10 ng of immunoprecipitated (IP) DNA was amplified using the Whole Genome Amplification kit (Sigma) according to the manufacturer's instructions. The mock ChIP-chip sample used preimmune antibody, rather than anti-Twist. For ChIP-seq, 50 ng of IP material was used to prepare a library (Johnson et al. 2007), and DNA sequencing of samples was performed by the Illumina protocol at Caltech Genome Center. The ChIP-seq input control was processed equivalently to the Twist ChIP-seq sample, except that it was not immunoprecipitated (no antibody or bead processing). Each ChIP-seq library was sequenced to a total of 9 million reads.

SELEX

SELEX experiments using in vitro binding to a column were carried out as described (Ogawa and Biggin 2011). See the Supplemental Text for more details, including processing of SELEX data.

Bioinformatics

ChIP-chip and ChIP-seq data processing: Methods used to call ChIP-chip versus ChIP-seq peaks are described in detail within the Supplemental Text. In brief, we used the ERANGE software suite to call peaks based on the number, orientation, and ratio of short sequenced reads relative to a background control. We considered an alternate peak caller (MACS), overlap of ChIP-seq regions with ChIP-chip regions, and the inclusion of known Twist targets to determine the threshold for calling Twist occupied sites (i.e.,

ChIP-seq signals). We selected a high confidence (HC) set of 513 sites based on high inclusion in ChIP-chip regions (87%), MACS regions (72%), and validated Twist targets (75%). We also selected a medium confidence (MC) set of 1099 regions based on the similarity in motif organization around these peaks (E-box, Fig. 3A).

ChIP-seq summit refinement

After ChIP-seq enriched regions were identified by the ERANGE program, post-processing was performed to refine the summit location by utilizing directional tag information. For each peak region, plus and minus tags were simultaneously shifted toward the imputed fragment center by a trial amount, ranging from 0 to 100 bp. The shift that maximized area overlap of the plus and minus tag density profiles (i.e., a measure of "directionality") was then implemented prior to calculating the location of the ChIP-seq tag count maximum ("summit").

Explanatory site interval

The interval for designating "explanatory sites" near ChIP-seq summits was estimated utilizing count statistics for the CACATG motif, due to its being the most prevalent E-box in the set of Twist regions. Specifically, the motif occurrences within increasing radii around peak centers (binned by 5 bp) were compared to the number expected from a Poisson distribution with the mean equal to the genome average density of CACATG motifs. When the probability of the observed number of counts coming from the Poisson model fell below 0.001, the distribution was deemed indistinguishable from random fluctuations, and the boundary of the previous bin was set to be the cutoff for explanatory sites ($\pm$50 bp from the summit).

Conservation analysis

Conservation at each base pair was assessed using phastCons scores (Siepel et al. 2005). Genome-wide scores for the fifteen-way insect alignment including *D. melanogaster, D. simulans, D. sechellia, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura, D. persimilis, D. willistoni, D. virilis, D. mojavensis, D. grimshawi, A. gambiae, A. mellifera*, and *T. castaneum* were downloaded from the UCSC genome gateway. Statistical analysis of the data is described in the Supplemental Methods.

Annotations

Precomputed annotation files for exons and introns were downloaded from the FlyBase website, release 5.27 (Tweedie et al. 2009). Here, exons and introns are mutually exclusive. 5′ UTRs data are from S. Celniker.

Acknowledgments

We thank the Caltech Jacobs Genome Facility members I. Antoshechkin and L. Schaeffer for library building and DNA sequencing, as well as D. Trout, B. King, and H. Amrhein for primary sequence data processing and visualization. We are grateful to A. Mortazavi and A. Kirilusha (Caltech Biology) for software and discussion of analysis; M. Biggin and S. Celniker (Lawrence Berkeley Lab) for sharing unpublished data; and M. Levine (University of California at Berkeley) for antibodies. K.I.F.-A. was funded by a NSF pre-doctoral fellowship, and S.P. was funded by The Gordon and Betty Moore Foundation. Work at Lawrence Berkeley National Laboratory was conducted under Department of Energy contract DE-AC02-05CH11231. This work was funded by the Functional Genomics Resource Center of the Caltech Beckman Institute, NIH grant R01GM077668 (A.S.), NIH grant U54HG004576 (B.J.W.), and the Bren Chair (B.J.W).

References

Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem* **94:** 890–898.

Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, Lefrancois P, Struhl K, Gerstein M, Snyder M. 2009. Mapping accessible chromatin regions using Sono-Seq. *Proc Natl Acad Sci* **106:** 14926–14931.

Bailey TL, Williams N, Misleh C, Li WW. 2006. MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* **34:** W369–373 (Web Server issue).

Bergman CM, Pfeiffer BD, Rincon-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, et al. 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biol* **3:** RESEARCH0086. doi: 10.1186/gb-2002-3-12-research0086.

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB. 2002. Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc Natl Acad Sci* **99:** 757–762.

Cao Y, Yao Z, Sarkar D, Lawrence M, Sanchez GJ, Parker MH, MacQuarrie KL, Davison J, Morgan MT, Ruzzo WL, et al. 2010. Genome-wide MyoD binding in skeletal muscle cells: A potential for broad cellular reprogramming. *Dev Cell* **18:** 662–674.

Castanon I, Von Stetina S, Kass J, Baylies MK. 2001. Dimerization partners determine the activity of the Twist bHLH protein during *Drosophila* mesoderm development. *Development* **128:** 3145–3159.

Chopra VS, Levine M. 2009. Combinatorial patterning mechanisms in the *Drosophila* embryo. *Brief Funct Genomics Proteomics* **8:** 243–249.

Crocker J, Tamori Y, Erives A. 2008. Evolution acts on enhancer organization to fine-tune gradient threshold readouts. *PLoS Biol* **6:** e263. doi: 10.1371/journal.pbio.0060263.

Davidson EH. 2006. *The regulatory genome: Gene regulatory networks in development and evolution*. Academic, Burlington, MA.

Erives A, Levine M. 2004. Coordinate enhancers share common organizational features in the *Drosophila* genome. *Proc Natl Acad Sci* **101:** 3851–3856.

Fullwood MJ, Ruan Y. 2009. ChIP-based methods for the identification of long-range chromatin interactions. *J Cell Biochem* **107:** 30–39.

Gonzalez-Crespo S, Levine M. 1993. Interactions between *dorsal* and helix-loop-helix proteins initiate the differentiation of the embryonic mesoderm and neuroectoderm in *Drosophila*. *Genes Dev* **7:** 1703–1713.

Grove CA, De Masi F, Barrasa MI, Newburger DE, Alkema MJ, Bulyk ML, Walhout AJ. 2009. A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell* **138:** 314–327.

Gurudatta BV, Corces VG. 2009. Chromatin insulators: Lessons from the fly. *Brief Funct Genomics Proteomics* **8:** 276–282.

Ho MC, Johnsen H, Goetz SE, Schiller BJ, Bae E, Tran DA, Shur AS, Allen JM, Rau C, Bender W, et al. 2009. Functional evolution of *cis*-regulatory modules at a homeotic gene in *Drosophila*. *PLoS Genet* **5:** e1000709. doi: 10.1371/journal.pgen.1000709.

Ip YT, Levine M, Small SJ. 1992a. The bicoid and dorsal morphogens use a similar strategy to make stripes in the *Drosophila* embryo. *J Cell Sci Suppl* **16:** 33–38.

Ip YT, Park RE, Kosman D, Bier E, Levine M. 1992b. The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive neuroectoderm of the *Drosophila* embryo. *Genes Dev* **6:** 1728–1739.

Ip YT, Park RE, Kosman D, Yazdanbakhsh K, Levine M. 1992c. *dorsal–twist* interactions establish *snail* expression in the presumptive mesoderm of the *Drosophila* embryo. *Genes Dev* **6:** 1518–1530.

Iyer V, Struhl K. 1995. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J* **14:** 2570–2579.

Jiang J, Rushlow CA, Zhou Q, Small S, Levine M. 1992. Individual dorsal morphogen binding sites mediate activation and repression in the *Drosophila* embryo. *EMBO J* **11:** 3147–3154.

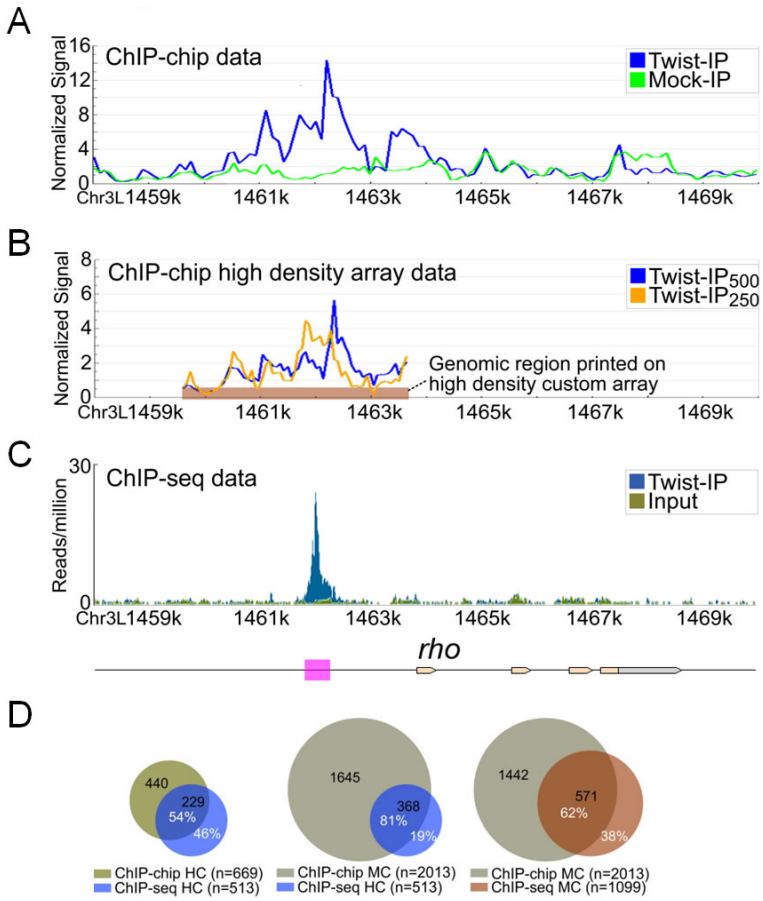
Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497–1502.

Landolin JM, Johnson DS, Trinklein ND, Aldred SF, Medina C, Shulha H, Weng Z, Myers RM. 2010. Sequence features that drive human promoter function and tissue specificity. *Genome Res* **20:** 890–898.

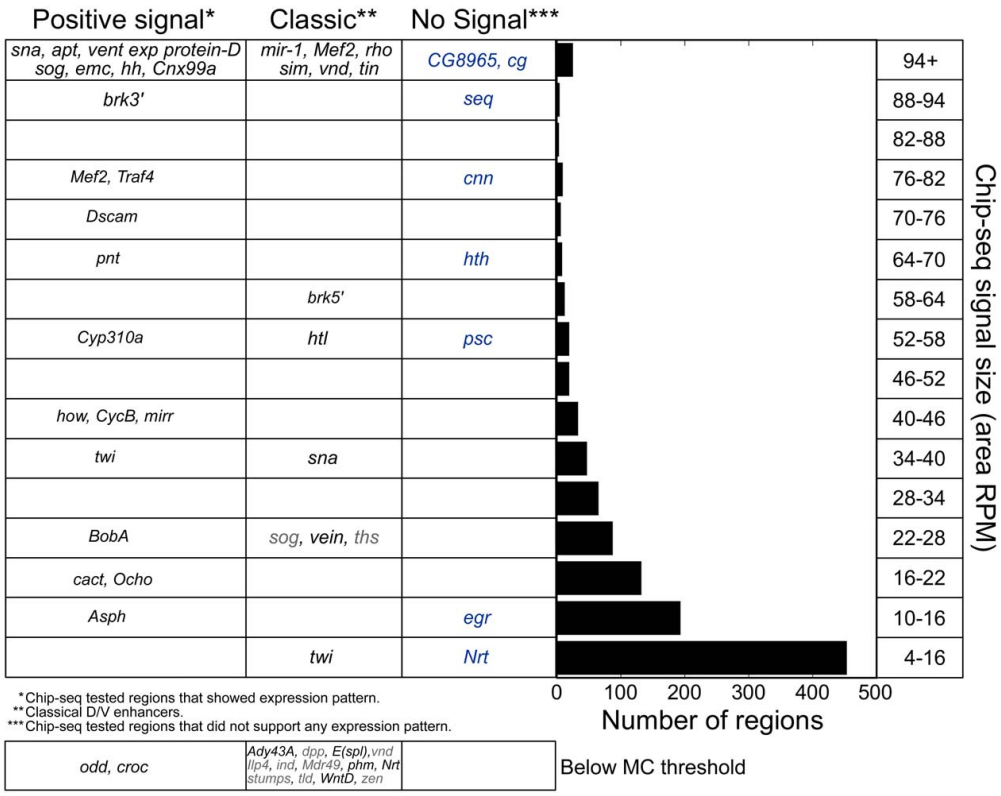
Lehmann M. 2004. Anything else but GAGA: A nonhistone protein complex reshapes chromatin structure. *Trends Genet* **20:** 15–22.

Levine M, Tjian R. 2003. Transcription regulation and animal diversity. *Nature* **424:** 147–151.
Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL, et al. 2008. Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol* **6:** e27. doi: 10.1371/journal.pbio.0060027.
Liberman LM, Stathopoulos A. 2009. Design flexibility in *cis*-regulatory control of gene expression: Synthetic and comparative evidence. *Dev Biol* **327:** 578–589.
Lusk RW, Eisen MB. 2010. Evolutionary mirages: Selection on binding site composition creates the illusion of conserved grammars in *Drosophila* enhancers. *PLoS Genet* **6:** e1000829. doi: 10.1371/journal.pgen.1000829.
MacArthur S, Li XY, Li J, Brown JB, Chu HC, Zeng L, Grondona BP, Hechmer A, Simirenko L, Keranen SV, et al. 2009. Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol* **10:** R80. doi: 10.1186/gb-2009-10-7-r80.
Markstein M, Markstein P, Markstein V, Levine MS. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc Natl Acad Sci* **99:** 763–768.
Markstein M, Zinzen R, Markstein P, Yee KP, Erives A, Stathopoulos A, Levine M. 2004. A regulatory code for neurogenic gene expression in the *Drosophila* embryo. *Development* **131:** 2387–2394.
Massari ME, Murre C. 2000. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Mol Cell Biol* **20:** 429–440.
Murre C, McCaw PS, Vaessin H, Caudy M, Jan LY, Jan YN, Cabrera CV, Buskin JN, Hauschka SD, Lassar AB, et al. 1989. Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. *Cell* **58:** 537–544.
Nam J, Dong P, Tarpine R, Istrail S, Davidson EH. 2010. Functional *cis*-regulatory genomics for systems biology. *Proc Natl Acad Sci* **107:** 3930–3935.
Ogawa N, Biggin MD. 2011. High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro. In *Methods in molecular biology* (ed. B Deplanke), Humana Press, Clifton, New Jersey (in press).
Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, Weng Z. 2007. Nucleosome positioning signals in genomic DNA. *Genome Res* **17:** 1170–1177.
Pepke S, Wold B, Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6 (11 Suppl):** S22–S32.
Reeves GT, Stathopoulos A. 2009. Graded dorsal and differential gene regulation in the *Drosophila* embryo. *Cold Spring Harb Perspect Biol* **1:** a000836. doi: 10.1101/cshperspect.a000836.
Sandmann T, Jakobsen JS, Furlong EE. 2006. ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in *Drosophila melanogaster* embryos. *Nat Protoc* **1:** 2839–2855.
Sandmann T, Girardot C, Brehme M, Tongprasit W, Stolc V, Furlong EE. 2007. A core transcriptional network for early mesoderm development in *Drosophila melanogaster*. *Genes Dev* **21:** 436–449.
Schuettengruber B, Cavalli G. 2009. Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. *Development* **136:** 3531–3542.
Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050.
Small S, Blair A, Levine M. 1992. Regulation of even-skipped stripe 2 in the *Drosophila* embryo. *EMBO J* **11:** 4047–4057.
Sosinsky A, Honig B, Mann RS, Califano A. 2007. Discovering transcriptional regulatory regions in *Drosophila* by a nonalignment method for phylogenetic footprinting. *Proc Natl Acad Sci* **104:** 6305–6310.
Stathopoulos A, Levine M. 2002. Linear signaling in the Toll-Dorsal pathway of *Drosophila*: Activated Pelle kinase specifies all threshold outputs of gene expression while the bHLH protein Twist specifies a subset. *Development* **129:** 3411–3419.
Stathopoulos A, Levine M. 2005. Genomic regulatory networks and animal development. *Dev Cell* **9:** 449–462.
Stathopoulos A, Van Drenth M, Erives A, Markstein M, Levine M. 2002. Whole-genome analysis of dorsal-ventral patterning in the *Drosophila* embryo. *Cell* **111:** 687–701.
Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R, et al. 2009. FlyBase: Enhancing *Drosophila* gene ontology annotations. *Nucleic Acids Res* **37:** D555–D559.
Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* **5:** 829–834.
Zeitlinger J, Zinzen RP, Stark A, Kellis M, Zhang H, Young RA, Levine M. 2007. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the *Drosophila* embryo. *Genes Dev* **21:** 385–390.
Zinzen R, Senger K, Levine M, Papatsenko D. 2006. Computational models for neurogenic gene expression in the *Drosophila* embryo. *Curr Biol* **16:** 1358–1365.
Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EE. 2009. Combinatorial binding predicts spatio-temporal *cis*-regulatory activity. *Nature* **462:** 65–70.

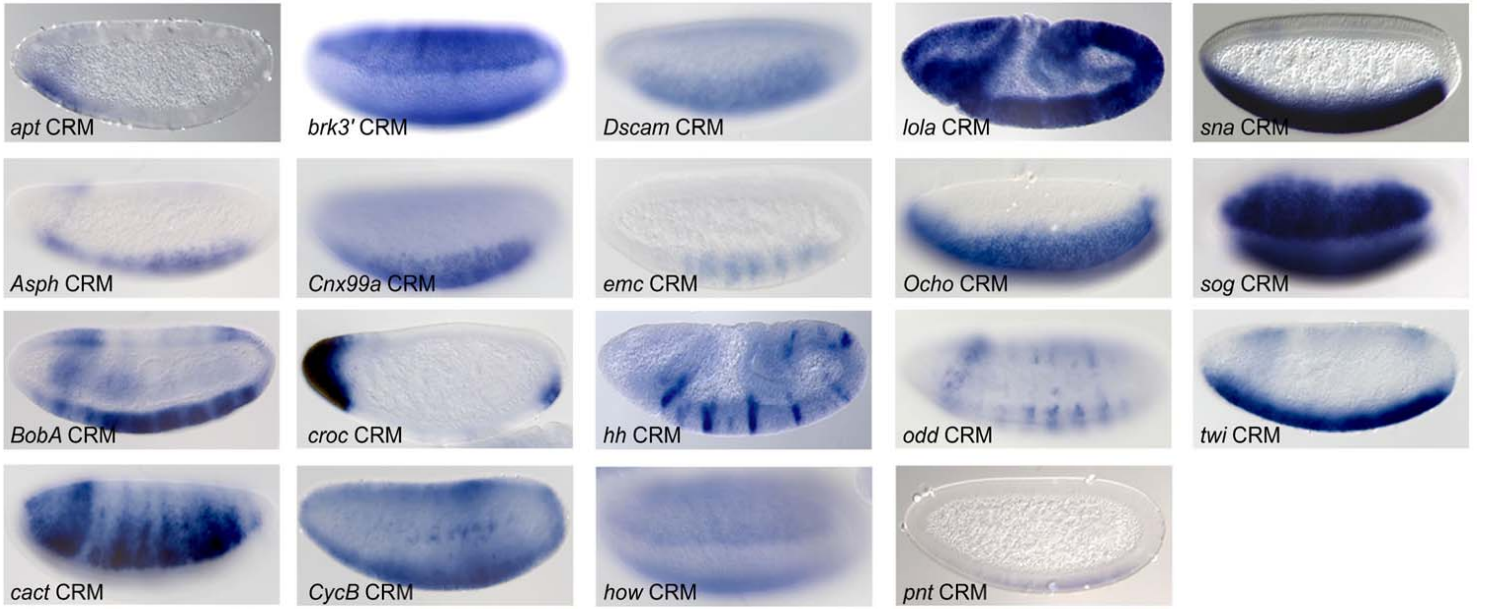
Received September 17, 2010; accepted in revised form January 4, 2011.



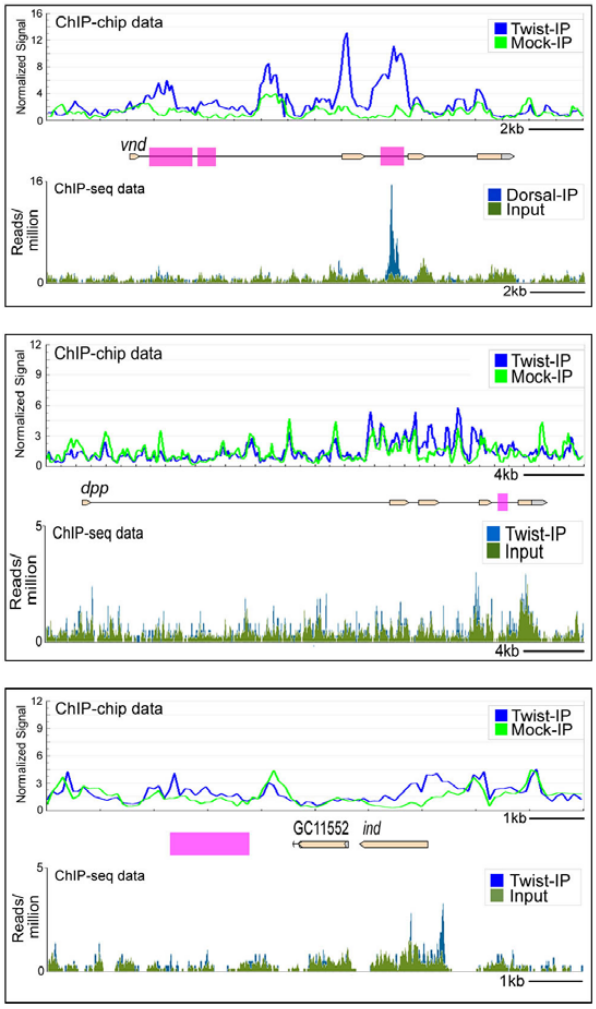
Supplemental Figure 1. *In vivo* **Twist occupancy determined by ChIP-Seq versus ChIP-chip and the isolation of CRMs.** (A) Twist ChIP-chip binding to a standard Nimblegen array at a representative locus, *rho*, relative to previously characterized early embryonic enhancer (pink box; Ip et al. 1992). (B) Twist ChIP-chip binding to a high-density custom array to same region for same Twist-IP (blue line) as used in (A); differences can be attributed to the assay method and data processing, rather than to the input chromatin lengths or other biological variation. Another independent Twist-IP prepared from smaller chromatin (sheared to ~250bp average) is shown in orange. Brown bar: location of the tiled regions on the custom array. (C) Twist ChIP-Seq-defined occupancy obtained using Twist antibody (blue) compared with sequenced input control DNA (green). (D) Venn diagrams showing the overlap between ChIP-chip and ChIP-Seq datasets of various sizes/FDRs. False Discovery Rate (FDR) of ~1% supported calling 513 high confidence (HC) ChIP-Seq regions and 669 HC ChIP-chip regions. FDR of 17% supported calling 1099 MC ChIP-Seq regions and 2013 MC ChIP-chip regions.



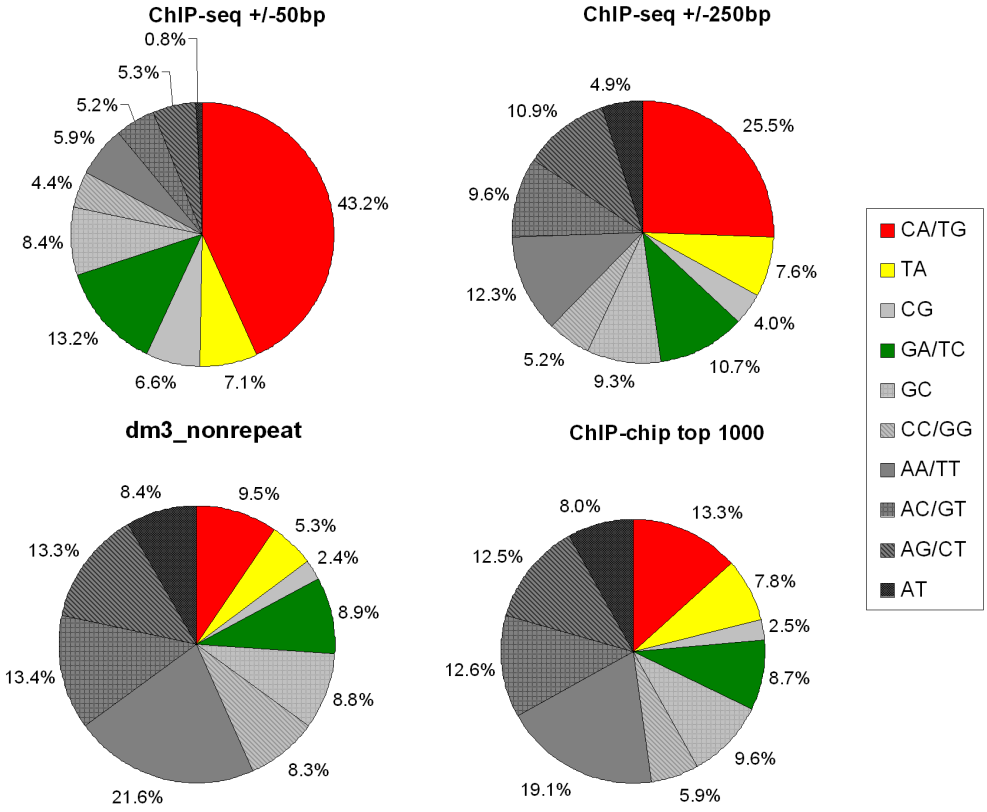
Supplemental Figure 2: Twist ChIP-Seq signals at known and candidate CRMs from prior studies. The number of Twist regions is shown ranked by signal size (reads per million in the entire area under the peak). As expected, lower ChIP signal regions are much more numerous than high signal regions. Regulatory regions that have previously been shown to support dorsal-ventral expression in the early embryo correspond to both large and small Twist ChIP-Seq peaks. In addition, regions that have been shown in this study to support expression and regions that failed to do so are distributed over the range of ChIP signal sizes.



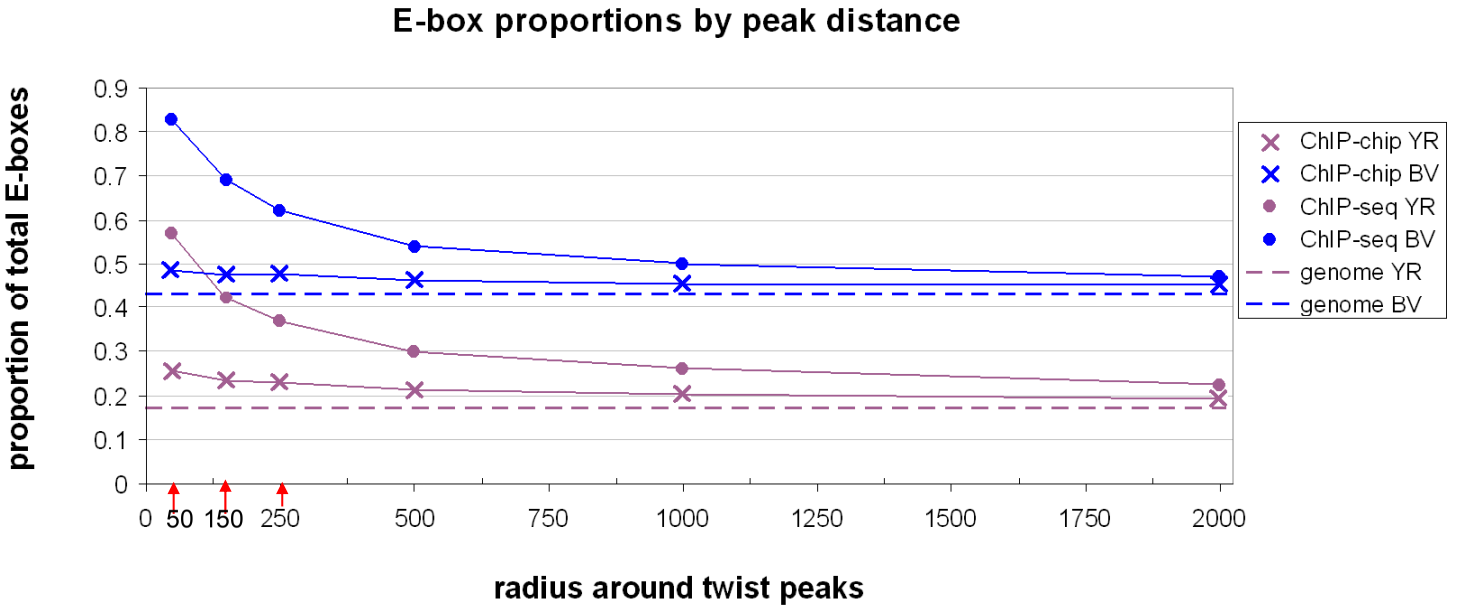
Supplemental Figure 3: Functional analysis of Twist regions by reporter gene assay. Twist regions were tested for their ability to support gene expression in a standard reporter gene assay using either *lacZ* or *cherry* reporter genes. *In situ* hybridization using riboprobes to *lacZ* or *cherry* were used to monitor gene expression supported by these DNA sequences in early embryos. Shown are the 19 of 31 tested regions found to support expression. Closest associated genes are indicated in the bottom corner of each panel; see Table 2a for exact coordinates of the DNA regions tested. Four additional regions found to support expression are shown in Fig. 1, for a total of 23 positives of 31 regions assayed.



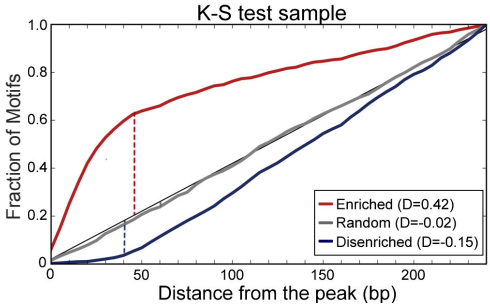
Supplemental Figure 4: Expression activity is not predicted by ChIP-Seq signal size. ChIP-chip and ChIP-Seq Twist data from this study are shown on the top and bottom of each panel, respectively. Pink boxes mark the locations of previously characterized enhancers. Twist signal is detected at the previously characterized *vnd* early embryonic enhancer located in the second intron (Stathopoulos et al. 2002), which is consistent with the early 1-3 hr timepoint assayed in this study. We do not detect significant Twist signal at a second *vnd* candidate enhancer which was identified more recently by ChIP-chip analyses at a slightly later developmental timepoint (Zeitlinger et al., 2007); perhaps the enhancers in the first intron support later or weaker gene expression. In the cases of *dpp* and *ind*, the sites shown are candidate enhancers based on motif presence and/or ChIP-chip binding. We did not see significant signals at these sites. *dpp* and *ind* are expressed in dorsal and dorsal-lateral regions of the embryo, which are outside the spatial domain of most Twist expression. These therefore fall into the group of previously discussed Twist targets that we call "Type III" (see text).



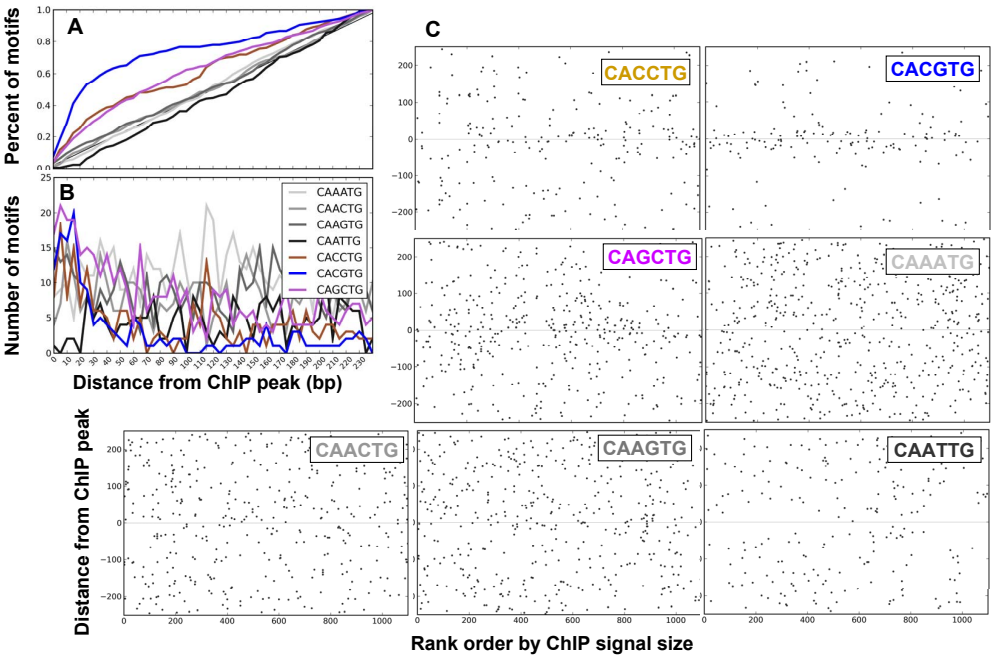
Supplemental Figure 5: Frequency of E-box instances in ChIP-Seq versus ChIP-chip close to the signal summit (±50bp) or at greater distance from it (±250bp). CANNTG E-boxes were tallied around Twist MC ChIP-Seq peaks, the largest 1,000 MC Twist ChIP-chip peaks, and the non-repeat fly genome. Displayed are the proportions of the different possible interior ten NN base pairs. When the areas very close (±50bp) to Twist ChIP-Seq peaks are compared to the wider ±250bp areas around Twist peaks, CA E-boxes predominate, suggesting that they dominate in supporting ChIP-detectable binding. There is also a distinct lack of AT E-boxes. The proportion of TA E-boxes remains relatively steady close to and farther from the peaks. The proportions of E-box cores around ChIP-chip summits are very similar to the genomic background distribution, suggesting that while ChIP-chip tiling arrays find larger domains putatively occupied by Twist, the peak of signal is far less accurate in identifying the explanatory Twist binding sites.



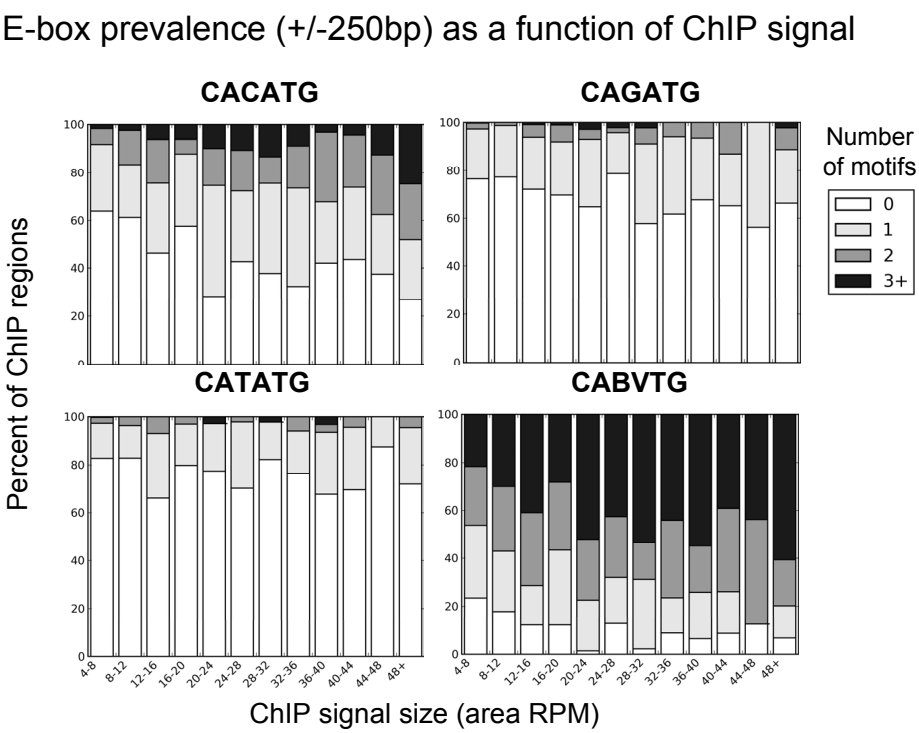
Supplemental Figure 6: Frequency of CAYRTG or CABVTG E-boxes within ChIP-chip or ChIP-Seq data as a function of distance from the summit. Twist 'explanatory' E-boxes were classed in two ways: the more canonical and stringent CAYRTG-core E-boxes (CA, TA, and CG) as well as the expanded CABVTG core suggested by our data (also including GA, GC, and CC). YR and BV E-boxes as a percent of all 10 possible E-boxes are shown in expanding radii out from the largest 1,000 Twist MC ChIP-chip peaks and the MC Twist ChIP-Seq peaks. They are compared to the distribution in the non-repeat genome. The ChIP-Seq data shows a marked enrichment of both types of explanatory E-boxes within ±50bp of ERANGE peaks (almost 85% of E-boxes are BV and almost 60% are YR) and this drops off exponentially with distance from the peak. The proportion of explanatory E-boxes is slightly greater near ChIP-chip summits as compared to the genomic background distribution.



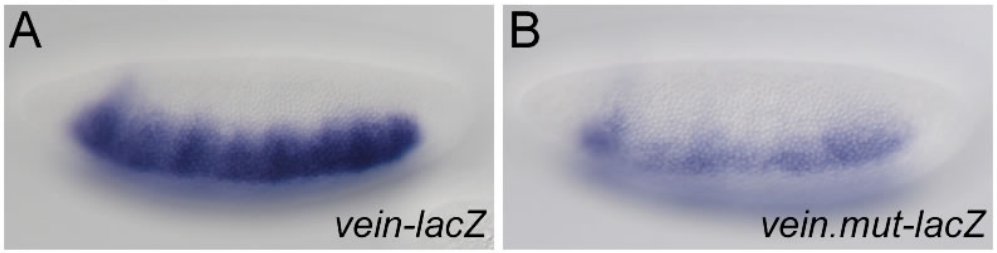
Supplemental Figure 7: Visual example of the K-S test. The Kolmogorov-Smirnov (K-S) test determines the degree of similarity between two distributions (see Supplemental Methods). In order to determine whether certain motifs were enriched or depleted relative to Twist peaks, their cumulative distributions (red, blue, and grey plots) were compared to the cumulative distribution function of a uniform distribution (black diagonal line). D (dotted vertical line) is the maximum distance between the motif distribution function and the uniform distribution function. While the P-value determines if a distribution is statistically the same as uniform instead of enriched or depleted, the absolute value of D reflects the spatial degree (bp around Twist peaks) of the enrichment or depletion of a motif. A large D absolute value reflects a large degree of enrichment/depletion; enriched motifs have positive D values and depleted motifs have negative D values. P-values reported are in base 10 (i.e. 2.2E-16 means 2.2*10-16)



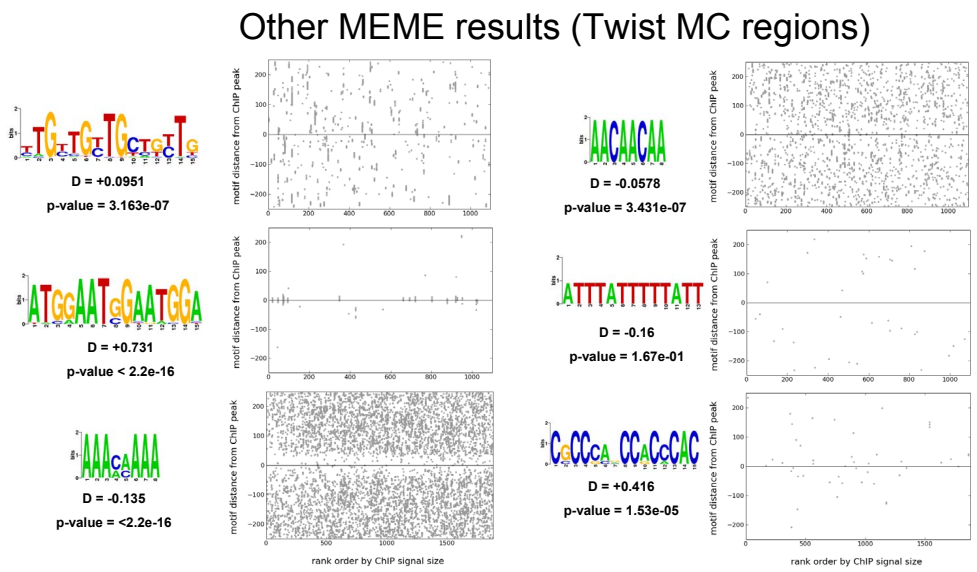
Supplemental Figure 8: Distribution of additional E-boxes within Twist ChIP-Seq data. The three CABVTG E-boxes not shown in Figure 4: (CACCTG, CACGTG, and CAGCTG) also show some enrichment relative to the peak. Of these, CAGCTG is the most prevalent. CACGTG (the third member of the CAYRTG E-boxes) occurs less frequently but is quite enriched around Twist peaks. The 4 CAANTG E-boxes are not enriched relative to Twist peaks, and in fact, the CAATTG palindrome is weakly depleted. See Supplemental Table 3 for the K-S values.



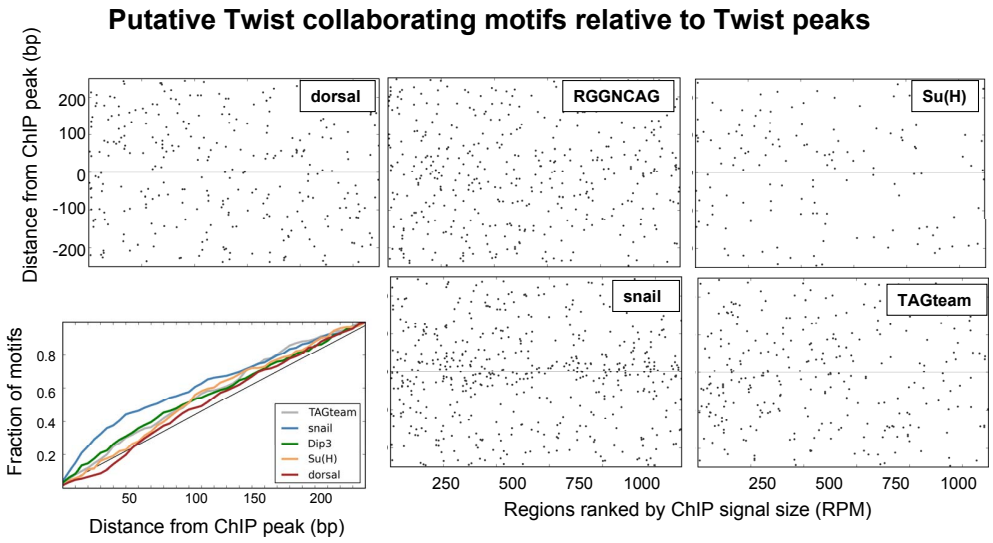
Supplemental Figure 9: E-box motif occurrence as a function of Twist ChIP-Seq signal size. The number of CACATG, CAGATG, CATATG, and CABVTG E-boxes were counted in a ±250bp radius around each Twist peak. MC Twist regions were ranked according to size (area RPM), and the percentage of regions containing 0, 1, 2, or 3 and more motifs is shown for each size category. CACATG motifs occur within about 50% of the whole MC dataset, but the larger peaks are more likely to have multiple occurrences of E-boxes. This trend does not hold true for CATATG and CAGATG, which occur in only about 25% of the peaks, and are most likely to occur singly. Viewed collectively, CABVTG E-boxes are present in the large ±250bp radius around over 90% of Twist peaks and are also more likely to occur multiply near large Twist peaks. This suggests that the largest signal size features are most likely to be driven by multiple binding sites.



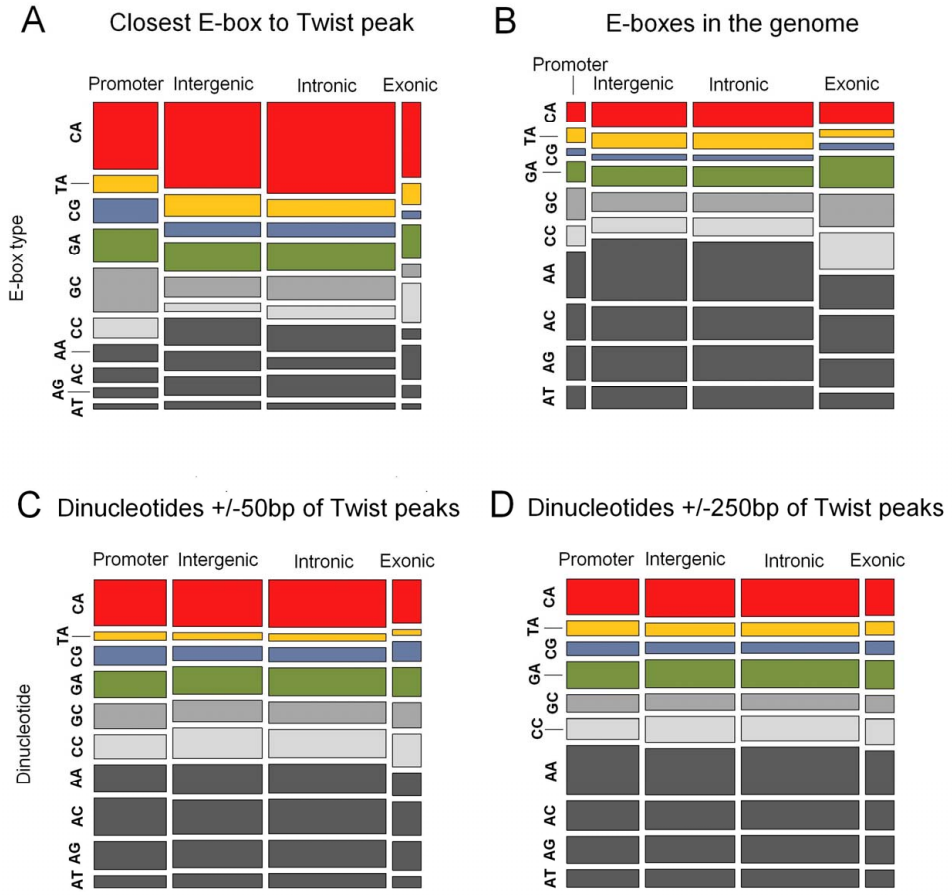
Supplemental Figure 10: *vein* CRM mutagenesis demonstrates the requirement for the explanatory E-box. We introduced a single base pair change within potential explanatory sites (CACATG > GACATG) we had defined within the *vein* CRM (A), characterized previously (Markstein et al. 2004). Mutating the explanatory CA-core E-box in this manner resulted in a dramatic loss of reporter gene expression (B). Reporter gene expression was abrogated such that the expression domain collapsed from 10-12 cells in width to 4-7 cells for the vein CRM; this effect is comparable to the expression of *vein* gene in *twist* mutant embryos (data not shown). Previously, the orientation of this same E-box was also shown to be important for *vein* CRM expression (Zinzen et al. 2006).



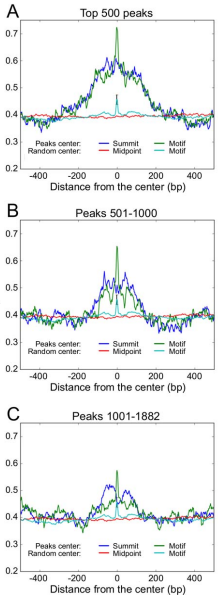
Supplemental Figure 11. MEME outputs. The other MEME outputs not shown in Figure 6 are displayed here and mapped back onto Twist MC regions at 85% threshold. Their K-S values are shown in Supplemental Table 3 where, from top to bottom by column, they are called MEME MC ±50 motifs 3, 4, 6, 7, 9, and 10.



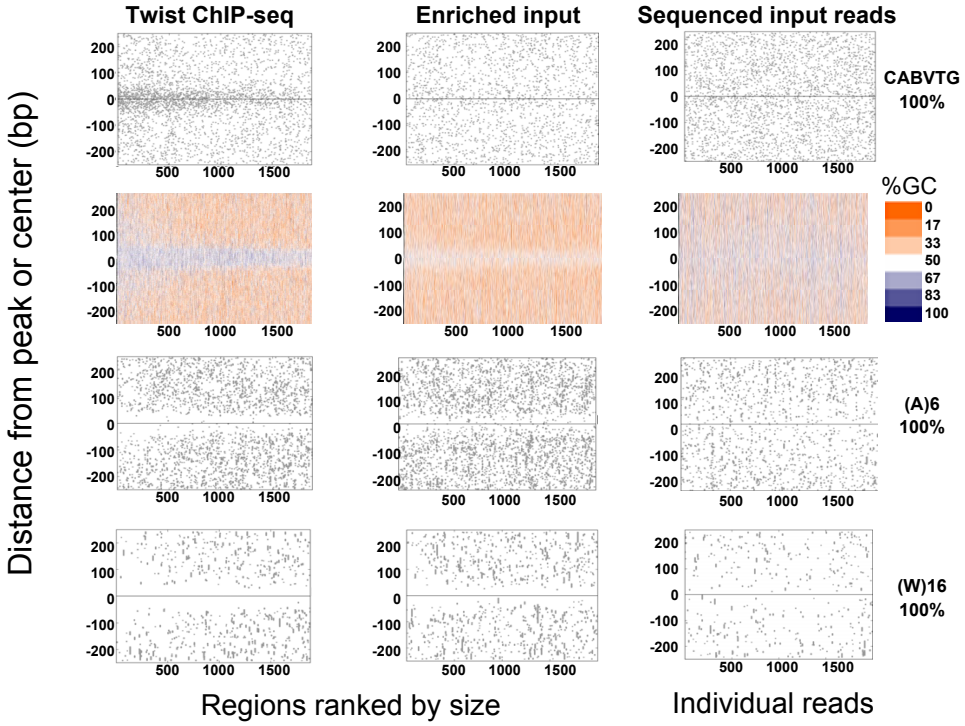
Supplemental Figure 12. Distributions of binding motifs for factors thought to interact with Twist. The motifs for Dorsal (SELEX – GGG(W3-5)CYV, 100% match) (Markstein et al. 2002; Zinzen et al. 2006; Liberman and Stathopoulos 2009); Zelda (TAGteam – YAGGYAG, 100% match) (ten Bosch et al. 2006); Suppressor of Hairless [Su(H) – BRTGRGAAH 90% match] (Bailey and Posakony 1995); RGGNCAG/Unknown (RGGNCAG, 100% match) (Stathopoulos et al. 2002); and Snail (RCARGWBB, 90% match) (Stathopoulos and Levine 2005) are shown relative to Twist peaks. If these factors interact directly with Twist to support expression through these predicted CRM regions, we would predict enrichment of the binding motifs relative to Twist peaks. The SELEX-derived Dorsal site [GGG(W3-5)CYV (A) as well as other previously described Dorsal sites (data not shown)] and Zelda are not enriched relative to Twist peaks. The Su(H) and RGGNCAG motifs are present and weakly clustered around the Twist peaks (B). Snail exhibits a significantly enriched binding site distribution near Twist summits, yet because the Snail consensus binding sequence overlaps with that of some Twist sites, the interpretation of this result with respect to probable Snail activity is not certain. See Supplemental Table 3 for the K-S values of these motifs.



Supplemental Figure 13. E-box and dinucleotide repeat frequencies under Twist ChIP-Seq peaks versus the genome. (A) Twist MC peaks (i.e. "shifted summits") were classed according to genomic location (as in Figure 6; see Supplementary Methods) and the closest E-boxes within ±50bp of Twist peaks in each category is shown. 23% of promoter proximal, 25% of intergenic, 23% of intronic, and 41% of exonic regions have no E-box within ±50bp. (B) The proportion of E-boxes in all genomic categories is shown. The proportion of CAGCTG E-boxes is greater in promoters than intergenic regions or introns, but it is still not as large as the proportion of CAGCTG E-boxes in Twist regions associated with promoters. In order to determine if the E-box proportions under Twist peaks is a direct result of dinucleotide frequencies in different regions of the genome, we analyzed all dinucleotides under the narrow ±50bp around Twist peaks (C) and the larger ±250bp radius (D). There is very little change in the frequency of dinucleotides under Twist peaks falling into different areas of the genome, suggesting that the proportional E-box difference between categories is not due to overall dinucleotide representation. There are slightly fewer A/T-rich dinucleotides very close to Twist peaks, which is consistent with an overall depletion of A/T-rich sequences near peaks (Supplemental. Figure 15).



Supplemental Figure 14. Conservation local to summits throughout peak rankings. The average PhastCon score is shown at every base pair around Twist-occupied sites ("peaks") and compared to average conservation distribution of 30 samples of 500 regions from the non-repeat dm3 genome ("random"). The "summit centered" plots are drawn relative to the shifted ERANGE peaks (Twist) and the "midpoint centered" plots are drawn relative to the centers of the randomly selected genomic background regions. The "motif centered" Twist plot was re-centered on the nearest CABVTG E-box (Twist explanatory motif) within ±150bp of the ERANGE summits, and regions with no such motif were left out. For the "motif centered" random plot, random regions were pre-screened to contain one of the CABVTG motifs. Relative to the genomic background, the entire area around Twist occupied sites is highly conserved in the HC sample (A). This occurs not just in the summits, but out to the broader area ±150bp. This conservation is even more increased when centering on the nearest CABVTG E-box, although the motif-centered random plot shows that CABVTG E-boxes in the *Drosophila* genome are preferentially conserved relative to the genomic background. The conservation of the 500 peaks added by dropping to the MC threshold is smaller overall (B), and the conservation of the additional 1,000 peaks from the LC threshold is even smaller (C). This may suggest that smaller peaks are less likely to be conserved or it may be a result of having more false positive peaks as the threshold is lowered.



Supplemental Figure 15. Distribution of motifs within the sequenced input DNA (i.e. sonicated chromatin). Twist ChIP-Seq regions are significantly depleted in highly A/T-rich sequences. This depletion is not specific to the ChIP because it is also observed for the input control chromatin library. Twist MC ChIP-Seq peaks are shown next to input control data of an equivalent number of regions (1099). See Supplemental methods for the origin of the different control samples shown. "Enriched input" contains regions selected as most significant from the input control over Twist. "Sequenced input reads" reads were randomly selected from all uniquely mapping reads in the input control. For Twist and enriched input, mapping is relative to the shifted summits. For the sequenced input reads, mapping is relative to the center of each 25bp read. Three motifs, the Twist explanatory E-box (CABVTG), AAAAAA [(A)6], and a string of any 16 A's or T's [(W)16] are shown for each dataset and compared to the overall GC content (averaged in 20bp windows).

Supplemental Table 2a:Novel Twist binding regions determined by ChIP-seq data

Associated gene name	Annotation ID	Tested Region Coordinates	Twist ChIP-seq signal size (RPM over whole region)	Supports Expression	Explanatory site	Defined by ChIP-chip/-seq	Peak Location
sna	CG3956	Chr2L:15485571-15487571*	233.4	Yes	CACATG	Yes/Yes	Exon[1]
apt	CG5393	Chr2R:19460067-19462066	229.6	Yes	CACATG	Yes/Yes	Intergenic
sog	CG9224	ChrX:1554000-15541901	160.7	Yes	CATCTG	Yes/Yes	Intergenic
ventally-expressed-protein-D	CG33200	Chr2R:18295567-18296766	145.6	Yes	CAGCTG	Yes/Yes	Promoter proximal
CG8965	CG8965	Chr2L:5870750-5871849	121	No	CACATG	Yes/Yes	Intron
emc	CG1007	Chr3L:742059-743659	119.5	Yes	CAGCTG	Yes/Yes	Intergenic
cg	CG8367	Chr2R:10061749-10062800	118	No	NA	Yes/Yes	Intron
hh	CG4637	Chr3R:18967451-18968550	117.1	Yes	CAGATG	Yes/Yes	Promoter proximal
Cnx99A	CG11958	Chr3R:25134141-25135377	114.8	Yes	CACATG	Yes/Yes	Intergenic
seq	CG32904	Chr2R:9076428-9077427	93	No	CACATG	Yes/Yes	Promoter proximal
brk	CG9653	ChrX:7214185-7215408	92.6	Yes	CATATG	Yes/Yes	Intergenic[2]
Mef2	CG1429	Chr2R:5847634-5848273	78	Yes	CATATG	Yes/Yes	Intergenic[3]
cnn	CG4832	Chr2R:9334816-9336990	77.6	No	CACATG	Yes/Yes	Promoter proximal
Traf4	CG3048	Chr2L:4369850-4370849	77.5	Yes	CACATG	Yes/Yes	Intergenic
Dscam	CG17800	Chr2R:3268100-3268475	73.9	Yes	CACATG	Yes/Yes	Intron
hth	CG17117	Chr3R:6427854-6428902	65.9	No	CACATG	Yes/Yes	Intron
pnt	CG17077	Chr3R:19169181-19170380	64.1	Yes	CACATG	Yes/Yes	Intron
Cyp310a	CG10391	Chr2L:18652293-18653292	53.2	Yes	CAGATG	Yes/Yes	Exon
psc	CG3886	Chr2R:8872127-8873594	53.1	No	CAGATG	Yes/Yes	Intergenic
how	CG10293	Chr3R:17870013-17871012	45	Yes	CATATG	Yes/Yes	Intron
CycB	CG3510	Chr2R:18690200-18691300	44.8	Yes	CACATG	Yes/Yes	Intergenic
mirr	CG10601	Chr3L:12670165-12671257	43.3	Yes	CACGTG	Yes/Yes	Intergenic
twi	CG2956	Chr2R:18936997-18938126	36.9	Yes	CACATG	Yes/Yes	Intergenic
BobA	CG12487	Chr3L:14945343-14947303	25.9	Yes	NA	Yes/Yes	Intergenic
cact	CG5848	Chr2L:16318750-16320750	21.8	Yes	CACATG	Yes/Yes	Intron
Ocho	CG3396	Chr3L:14968341-14970091	20.9	Yes	CACATG	Yes/Yes	Intergenic
egr	CG12919	Chr2R:5964200-5966200	13.6	No	CACGTG	Yes/No	Promoter proximal
Asph	CG8421	Chr2R:11997142-11999642	12.7	Yes	CACATG	Yes/No	Intron
odd	CG3851	Chr2L:3608800-3609500	<4	Yes	NA	Yes/No	Intergenic
croc	CG5069	Chr3L:21471259-21475259	<4	Yes	NA	Yes/No	Intergenic

Nrt	CG9704	Chr3L:16759474-16762173	<4	No	NA	Yes/No	Intergenic

* *D.melanogaster (*dm3, April 2006, BGDP release 5) genome assembly.

1 This *sna* enhancer is located in the 1st intron and 2nd exon of *Tim17b2*

2 This *brk* enhancer is located in the 5th intron of *Atg5*

3 This *Mef2* enhancer is located in the 2nd intron of *CG12130*

Supplemental Table 2b: Names and coordinates of classical D/V enhancers

Associated gene name	Annotation ID	Enhancer Region Coordinates	Twist ChIP-seq signal size (RPM over whole region)	Enhancer type I / II / III /Citation	Explanatory site	Defined by ChIP-chip/-seq	Peak Location
mir-1	CR32958	Chr2L:20480577-20481741	235.6	I / Zeitlinger et al. (2007)	CACATG	Yes/Yes	Promoter proximal
Mef2	CG1429	Chr2R:5819035-5820049	234.8	I / Nguyen and Xu (1998)	CACATG	Yes/Yes	Intron
rho	CG1004	Chr3L:1461823-1462121	231.6	II / Ip et al. (1992)	CA(CA/TA)TG	Yes/Yes	Gene distal
sim	CG7771	Chr3R:8895836-8896466	139.8	II / Kasai et al. (1998)	CATATG	Yes/Yes	Intron
vnd	CG6172	ChrX:485983-487688	139.7	II / Stathopoulos et al. (2002)	CACATG	Yes/Yes	Intron
tin	CG7895	Chr3R:17205671-17206053	106.8	I / Yin et al. (1997)	CACATG	Yes/Yes	Intron
brk	CG9653	ChrX:7190967-7191464	58.1	II / Markstein et al. (2004)	CACATG	Yes/Yes	Intergenic
htl	CG7223	Chr3R:13875600-13876391	52.8	I / Stathopoulos et al. (2004)	CACATG	Yes/Yes	Intron
CG4221	CG4221	Chr3R:11746513–11748343	39.4	I / Sandman et al. (2007)	NA	Yes/Yes	Exon
sna	CG3956	Chr2L:15478170-15481082	39	I / Ip et al. (1994)	NA	Yes/Yes	Intergenic
retn	CG5403	Chr2R:19145341–19147650	38.5	I / Sandman et al. (2007)	CATCTG	Yes/Yes	Intron
sog	CG9224	ChrX:15518731-15519122	27.9	III / Markstein et al. (2004)	NA	Yes/Yes	Intron
vein	CG10491	Chr3L:5828771-5829267	24.1	II / Markstein et al. (2004)	CACATG	Yes/Yes	Intron
ths	CG12443	Chr2R:7681727-7682234	22.7	III / Stathopoulos et al. (2002)	CAGCTG	Yes/Yes	Intron
T48	CG5507	Chr3R:22707641–22709469	21	I / Sandman et al. (2007)	CAGCTG	Yes/Yes	Intron
cact	CG5848	Chr3R:22707641–22709469	16	I / Sandman et al. (2007)	CACATG	Yes/Yes	Intron
twi	CG2956	Chr2R:18932428-18933842	7.2	I / Thisse et al. (1992)	NA	Yes/Yes	Promoter proximal
Ady43A	CG1851	Chr2R:3133869-3134085	<4	II / Markstein et al. (2004)	NA	No/No	Promoter proximal
CG32372	CG32372	Chr3L:7495,261–7497085	<4	I / Sandman et al. (2007)	NA	No/No	Intron
CG8788	CG8788	Chr2R:4650254–4650882	<4	I / Sandman et al. (2007)	NA	Yes/No	Intron
crb	CG6383	chr3R:20123093–20123803	<4	I / Sandman et al. (2007)	NA	Yes/No	Intron
dpp	CG9885	Chr2L:2456346-2456884	<4	III / Huang et al. (1993)	NA	No/No	Intron
E(spl)	CG8365	Chr3R:21864777-21865879	<4	II / Papatsenko et al (2005)	NA	Yes/No	Promoter proximal
Ilp4	CG6736	Chr3L:9797449-9797743	<4	I / Stathopoulos et al. (2002)	NA	No/No	Intron
ind	CG11551	Chr3L:15032420-15033835	<4	III / Stathopoulos et al. (2005)	NA	No/No	Intergenic

Mrd49	CG3879	Chr2R:8833934-8834294	<4	I / Zeitlinger et al. (2007)	NA	No/No	Promoter proximal
phm	CG6578	Chr2R:19875348-19875790	<4	II / Markstein et al. (2004)	NA	Yes/No	Intron
stumps	CG31317	Chr3R:10423060-10424098	<4	I / Stathopoulos et al. (2004)	NA	Yes/No	Intron
tld	CG6868	Chr3R:20574723-20575519	<4	I / Kirov et al. (1994)	NA	No/No	Promoter proximal
vnd	CG6172	ChrX:477514-479251	<4	II / Zeitlinger et al. (2007)	NA	Yes/No	Intron
WntD	CG8458	Chr3R:9118955-9119462	<4	I / Sandman et al. (2007)	NA	Yes/No	Promoter proximal
zen	CG1046	Chr3R:2579866-2581378	<4	III / Markstein et al. (2004)	NA	Yes/No	Intergenic

Supplemental Table 3: Statistics showing enrichment or depletion of various motifs relative to Twist MC ChIP-Seq peaks

E-boxes	D (K-S test)	P-value (K-S test)	deg. freedom	P-value (T-test)
CAAATG	*0.051*	*5.63E-02*	*555*	*2.34E-01*
CAACTG	*0.043*	*2.26E-01*	*394*	*4.08E-01*
CAAGTG	*0.049*	*1.19E-01*	*444*	*1.49E-01*
CAATTG	*-0.077*	*7.95E-02*	*214*	*1.31E-01*
CACATG	0.426	< 2.20E-16	841	< 2.20E-16
CACCTG	0.216	9.28E-10	222	2.55E-08
CACGTG	0.440	< 2.20E-16	157	< 2.20E-16
CAGATG	0.290	< 2.20E-16	372	< 2.20E-16
CAGCTG	0.213	< 2.20E-16	402	< 2.20E-16
CATATG	0.191	1.03E-09	284	1.48E-09
CAYRTG	0.371	< 2.20E-16	1284	< 2.20E-16
CABVTG	0.310	< 2.20E-16	2283	< 2.20E-16
CANNTG - not CA/GA/TA	0.091	< 2.20E-16	2394	< 2.20E-16
CANNTG	0.188	< 2.20E-16	3894	< 2.20E-16
collaborators	**D (K-S test)**	**P-value (K-S test)**	**deg. freedom**	**P-value (T-test)**
RGGNCAG	0.112	3.20E-05	412	8.18E-05
su(H)	0.123	4.45E-04	136	2.61E-02
dorsal	*0.059*	*1.40E-01*	*284*	*1.91E-01*
TAGteam	*0.116*	*1.56E-02*	*287*	*5.15E-05*
snail	0.222	< 2.20E-16	513	< 2.20E-16
MEME Twist MC ± 50bp	**D (K-S test)**	**P-value (K-S test)**	**deg. freedom**	**P-value (T-test)**
MC +/-50 motif 1	0.467	< 2.20E-16	1393	< 2.20E-16
MC +/-50 motif 2	0.344	< 2.20E-16	1585	< 2.20E-16
MC +/-50 motif 3	0.095	3.16E-07	826	6.02E-09
MC +/-50 motif 4	0.731	< 2.20E-16	79	< 2.20E-16
MC +/-50 motif 5	0.228	< 2.20E-16	1066	< 2.20E-16
MC +/-50 motif 6	-0.135	< 2.20E-16	4450	< 2.20E-16
MC +/-50 motif 7	-0.058	3.43E-07	2223	2.21E-04
MC +/-50 motif 8	0.287	< 2.20E-16	312	< 2.20E-16
MC +/-50 motif 9	*-0.160*	*1.67E-01*	*34*	*2.73E-01*
MC +/-50 motif 10	0.416	1.53E-05	31	2.97E-05

grey italics = not significantly enriched or disenriched near Twist peaks

D > 0 refers to motifs that are enriched relative to Twist peaks; D < 0 to disenriched

P-values are in base 10 i.e. 2.2E-16 means 2.2*10^-16

MEME motifs 2, 8, 1, and 5 are shown in Fig. 5. The other MEME motifs are shown in Suppl. Fig.11

E-box motifs are shown in Fig. 3 and Suppl. Fig. 8. 'Collaborators' are shown in Suppl. Fig. 12

See Supplemental Methods for a description of the test statistics

Supplemental methods

ChIP-chip experimental design and processing. Arrays from standard catalog of Roche Nimblegen were used for this experiment covering the entire *Drosophila melanogaster* genome. The set of three arrays (385,000 probes/array) contain 50-mer probes spaced by 48 nucleotides on the genome. Each array was hybridized with two samples - genomic control DNA labeled with Cy3 and experimental sample labeled with Cy5. Two samples were hybridized to the arrays: Twist and mock sample as control (i.e. pre-immune). Each measurement was performed using a single biological replicate. The hybridizations were performed at a Nimblegen facility, and both the raw data and Cy5/Cy3 ratios for each array (Cy5=635 nm, Cy3=532 nm) were made available to us for analysis.

Design of custom array for ChIP-chip experiment. A custom array (Nimblegen 4-plex technology, 72,000 probes,) was designed to confirm the above results and also probe the neighborhoods of high-confidence transcription factors in more detail. Two sets of probes were included in the array: (i) Probes were tiled (60 mer probes, 5 nucleotide spacing) within 6 kb upstream and 1 kb downstream of ATG sites of 288 high-confidence transcription factors in *Drosophila melanogaster.* The list of transcription factors is available on request; (ii) Probes were also tiled (60 mer probes, 5nucleotide spacing) within 1 kb upstream and downstream of 1,600 peaks detected in the earlier ChIP-chip experiments. In total, the array contained 71,000 60-mer probes from the *D. melanogaster* genome and 1000 random sequences as control.

ChIP-chip bioinformatics. The data from all arrays were normalized using quantile normalization procedure. After normalization, ratios of Cy5/Cy3 were taken for each sample for further analysis. The original array design was based on V4 release of the Drosophila genome. Therefore, normalized data were mapped on to V5 genome assembly (dm3, April 2006) examined visually for validation.

ChIP-chip peak finding was conducted as previously described (MacArthur et al. 2009). First, quantile normalized data for each probe was replaced by the mean signal of all probes within +/-350 nucleotides from it. This smoothing step was performed in the logarithmic scale. All probes with normalized smoothed signal above 90th percentile in the array (normalized signal=2, high signal probes) were considered for further analysis. Multiple high signal neighboring probes (maximum gap 200 nucleotides) were combined into summits with height equal to the highest smoothed intensity within the region.

ChIP-Seq bioinformatics. Sequenced reads were trimmed to the first 25 base pairs and mapped onto the dm3 (April 2006, BGDP release 5) *Drosophila melanogaster* genome using bowtie 9.1 (Langmead et al., 2009). No more than two mismatches were allowed. Low-copy multireads (defined as reads mapping in 2 to 10 places) were allowed. Chromosomes U and the Het chromosomes were not used in the downstream analyses.

The ERANGE 3.1 software package was used to identify regions enriched in ChIP-Seq defined Twist occupancy. ERANGE finds areas in the genome that are densely occupied by reads and then identifies those that exceed signals in the background sample (sonicated input DNA) (Pepke et al. 2009). Regions that do not display proper left/right read directionality are discarded (see also main text). A custom code was used to computationally call a ChIP-Seq signal maximum location (the "shifted summit"), which introduced a shift in the position attributed as the "peak" based on the degree of read directionality. For simplicity, the shifted summit is reported as one nucleotide.

In order to get a broad view of what to expect based on the ChIP-Seq experimental assay as well as the bioinformatics assay, several different types of controls were used. For the genomic background, the dm3 genome was used minus UCSC simple and tandem repeats and minus the Chromosomes U and the Het chromosomes. In order to assay reads that could be sequenced, reads that mapped uniquely to the genome were selected at random ("sequenced control reads"). In order to determine which places in the genome were sequenced well ("aggregated control"), ERANGE was run on the sonicated input DNA library requiring only two reads per region (no directionality requirement was used and no enrichment relative to another library was required). In order to determine which places in the genome displayed proper read directionality and were overrepresented in the sequenced input control library relative to twist ("enriched control"), ERANGE was run on the input DNA library vs. twist, requiring at least a 1% enrichment per region in the input DNA and a minimum of two reads per region. The directionality filter was used as for Twist regions and the peaks were subsequently shifted using the same algorithm as for the Twist peaks.

A second independent ChIP-Seq algorithm and software package, MACS 1.3.5 (Zhang et al. 2008), was also used on the same Twist and input control datasets, and we report both sets of "peak calls" (Supplemental Table 4). The effective genome size used was 1.69e8, tag size 25, band width 300, model fold 7, and P-value cutoff 1e-5. There were no major discrepancies between motif occurrences relative to ERANGE and MACS calls nor to the respective MEME outputs (data not shown).

Selection of confidence thresholds. None of the distributions of ChIP signals, under

any algorithm, displayed a crisp natural discontinuity that would clearly define "occupied" versus "unoccupied" states. ERANGE was first run on ChIP-Seq data with a stringent gradient of parameters, and the different region sets were evaluated for sensitivity and specificity by their inclusion of (1) validated, functional Twist binding regions; (2) their overlap with an independent region calling algorithm, MACS and (3) the likelihood that the low-confidence end of the region sets were 'real' as judged by inspection of the read distribution in ChIP and background data. As a result, we set the ERANGE high-confidence (HC) signal and enrichment thresholds at 14 RPM minimum (reads in the region per million in the dataset), 1 RPM minimum peak height, and 3-fold enrichment over the control sample), resulting in 513 regions (false discovery rate (FDR) <1%, where the ERANGE FDR reflects the relative number of peaks called when using the same parameters to call the control library over the twist library). Medium confidence (1099 peaks) and lower-confidence (2000 peaks) were called with the same enrichment ratio and minimum peak height but instead using region RPM thresholds of 4 (FDR 17%) and 2 (FDR 83%), respectively. The MC threshold was selected because of the similarity of motif distributions around peaks compared to the HC regions (Figure 4A), and the LC threshold was selected primarily to demonstrate what happens when selecting a very low informatics threshold (shown in Figure 3A and Supplemental Figure 15).

For comparison sake, HC and MC sets of ChIP-chip regions were defined using equivalent FDR measures as found for ChIP-Seq. To this end, boundaries of ChIP-chip regions were defined using a threshold of 3.8 to identify 669 ChIP-chip regions (HC set; FDR<1%) and a threshold of 6 to identify 2013 ChIP-chip regions (MC set; FDR 17%). We report the MC region boundaries as well as the size and location of the "summit" of each region, defined as the midpoint of the highest part of each region (Supplemental Table 5).

As expected, the weaker ChIP-Seq signals are most numerous in their respective distributions (Supplemental Figure 2), which means that the computational threshold selected for inclusion has a large impact on subsequent VENN comparisons of Twist set membership. ChIP-chip processed data typically identified physically broader regions on the chromosome, partly because array processing algorithms require multiple positive tiles to make a signal call. Furthermore, the array data appear to compress the ChIP signal range compared with ChIP-Seq, bringing the strongest signal closer to the weakest one in the distribution and this, along with other technical differences, may account for the decrease in overlap observed when the HC ChIP-Seq set is compared with MC versus HC ChIP-chip sets (81% versus 54%).

Acquisition of SELEX data and processing. SELEX was performed according to a

previously published method (Roulet et al. 2002) and a standard SAGE protocol (http://www.sagenet.org/protocol/index.htm) with some exceptions, as follows (for further details see Ogawa 2011). 72 bp DNA oligoes were synthesized with three different end pairs each containing a restriction enzyme site (*Bam*HI, *Bgl*II, or *Hin*dIII) and 20 bp priming sequences for PCR amplification:
Random72: GGATTTGCTGGTGCAGTACAGT-GGATCC-$(N)_{16}$-GGATCC-TTAGGAGCTTGAAATCGAGCAG
Random72R: TCCATCGCTTCTGTATGACGCA-AGATCT-$(N)_{16}$-AGATCT-GTCCTAACCGACTCCGTTGATT
Random72HR: TCCATCGCTTCTGTATGACGCA-AAGCTT-$(N)_{16}$-AAGCTT-GTCCTAACCGACTCCGTTGATT

His-tagged Twist protein was bound to TALON Metal Affinity Resins (Clontech). For the first round of SELEX, 10 ng of the random 72 bp ds DNA oligonucleotides was incubated with the protein bound resin. The input DNA for subsequent rounds of SELEX was derived by PCR amplifying $1/10^{th}$ of the DNA eluted from the previous round.

For all rounds, SELEX-bound DNA was amplified by PCR according to SAGE protocol and then digested with the appropriate restriction enzyme to isolate the 22 bp fragment which includes the Twist-binding sequence. Approximately 1 µg of the 22 bp DNA fragments were ligated to make concatamers in a 10 µl volume at 16°C overnight. The concatemer DNA was treated with T4 DNA polymerase (NEB) and DNA polymerase I Klenow fragment (NEB) with dNTP mixture at room temperature for 30 min. After heat-inactivation at 65°C for 5 min, the DNA was separated by 2% agarose (Invitrogen, UltraPure agarose) gel electrophoresis. DNA of 300 to 1000 bp was isolated from the gel and purified by using QIAquick Gel purification kit (Qiagen). The resulting concatemer DNA was ligated with *Sma*I-digested pUC19 plasmid, and subsequently the ligation mixture was used to transform DH10B E. coli (Invitrogen ElectroMAX cells). Plasmid DNAs from more than 96 clones were sequenced to obtain sequences of over 1,000 individual DNAs. The data presented are 17 bp reads, on average (Supplemental Table 6).

Two SELEX experiments were performed to analyze the binding preference for Twist. Each involved 5 rounds of amplifications for a total of 10 total datasets. For experiment one, rounds 4 and 5 were sequenced; for experiment two, rounds 2,3, and 4 were sequenced. The data for these 5 rounds were pooled, and the number of E-boxes in the entire dataset was counted (Figure 2). MEME was run on the SELEX sequences, and in addition to the CATATG/CACATG E-box, an –AYRTG sequence (suggesting a partial E-box) was also returned (data not shown). E-boxes are present in approximately 50% of the SELEX sequences and of the

remaining 50%, the majority contain a partial (5-mer) E-box. This may be due to the enzyme cut sites and sequencing or possibly to Twist binding a partial E-box. We see no such representation of the partial E-boxes at ChIP-Seq *in vitro* peaks.

MEME analysis. MEME was run on the MC Twist ChIP-Seq ERANGE regions ±50 bp from the peaks (i.e. "shifted summits") in order to capture the pieces of DNA that show the highest enrichment of explanatory E-boxes (Figure 2, Figure 3, Supplemental Figure 8). MEME 3.0.8 was used, using the "zoops" model, 6 bp minimum, and 15 bp maximum motif widths. MEME finds sequences that are similar to each other but statistically unlikely to be found in the local background of the sample (Bailey et al. 2006). The MEME results were mapped at 85% match to the output PSFM's onto the parent set of Twist regions or the control datasets (Figure 5, Supplemental Figure 11).

Motif mapping. Scatter plots were made in order to visualize the distribution of motifs relative to Twist peaks (i.e. "shifted summits"). Motifs were mapped on to the genome, and each dot on a scatter plot reflects the distance between the center of the motif and its respective Twist peak. Negative values are to the left of the peaks in the reference genome, and positive numbers are to the right.

Density plots (i.e. Figure 3B, top panel) were made by taking the absolute distance of each motif from its peak and then summing for the entire dataset the number of motifs in 5 bp windows outward from the peaks. Cumulative density plots (i.e. Figure 3B, bottom panel, Supplemental Figure 7) are another way of reporting the data in the density plots, where the cumulative fraction of the motifs represented in each 5bp window in (from 0 total motifs found at the peak to 100% of the motifs encountered at the maximum 250 bp distance from the peak).

A Kolmogorov-Smirnov (K-S) statistical test was performed to determine whether motifs were enriched, depleted or uniformly distributed relative to the set of Twist peaks. This method tests the null hypothesis that a distribution of motif distances relative to Twist peaks is distributed uniformly. Distributions of these distances for motifs that are unrelated to binding are expected to be statistically similar to the uniform distribution; those that are related to binding are expected to be different from uniform. The statistic for testing these hypotheses is the maximum distance between the empirical cumulative distribution function of the distances between motifs and peaks and the cumulative distribution function of a uniform. This distance is known as the "D" value (D values and both types of distributions are illustrated in Supplemental Figure 7). Thus we can obtain P-values for the probability of the null hypothesis and reject the null hypothesis when the P-value is too small. All regions were made equal length (±250 bp around each peak) for these tests. A small P-value (threshold $1*10^{-3}$) means that a motif

distribution is not significantly different from uniform and is instead enriched or depleted relative to Twist peaks.

To relate the K-S test results to a more familiar statistic, we also performed a Student's T-test. The T-test is used here to test whether the mean of the observed motif distance from the peak is equal to the mean of the assumed uniform distribution on the standardized regions. Since we standardized the maximum distance from the peak to 250 bp, the mean is 125 bp, and so the T statistic reports whether the mean of each motif is different from 125 bp. Note that it is possible to have a distribution quite different in shape from the uniform distribution and still have the same mean. The K-S test would determine that the two are significantly different while the T-test would not. In this sense, the K-S test is more powerful than the T-test. In any case, the statistical conclusions from the T-test and the K-S test agree for our observed distributions (see the P-values for both tests in Supplemental Table 3). P-values reported are in base 10 (i.e. 2.2E-16 means $2.2*10^{-16}$)

Genome location analysis. The gene models we used were primarily based on published FlyBase introns and exons but were additionally informed by a set of promoters active in the embryo (generously provided by S. Celniker). We used these data to class the genome into four mutually exclusive categories. "Promoter proximal" refers to any summit that occurs within a Celniker promoter or 500 bp upstream. "Exonic" refers to any FlyBase exon excluding any regions that fall into the promoter proximal category. "Intronic" regions are any regions within the gene body (from FlyBase TSS or Celniker promoter, whichever is upstream, to the last exon) that are not in the exonic or promoter proximal categories. Intergenic regions are outside of gene bodies and had repeats (from UCSC tandem repeats and repeat masker) removed.

In order to accurately represent the nature of the ChIP-Seq input control data, we used it in three different ways. "Random sequenced input reads" is a set of reads from the input control that map uniquely to the genome. It represents the areas of the non-repeat genome which are able to be sonicated and sequenced. "Aggregated input control" regions were created by allowing ERANGE to run on the input control without a directionality filter or an enrichment requirement. These regions represent places in the genome that have an aggregation of input reads but no other requirements that the reads behave similarly to ChIP-Seq peaks. The "enriched input control" contains regions where the input control library is enriched over Twist and also displays the same left/right read directionality required for Twist (see also main text) .

The number of ChIP-chip and control regions in each dataset was picked to be the same number as MC Twist regions. We chose the largest ChIP-chip and

aggregated control regions (by area), the enriched control regions that were most highly enriched over Twist, and a random sample of sequenced control reads. In order to assign regions to each genomic category, we used the shifted summits of Twist ChIP-Seq and enriched control regions, the highest point of the aggregated control regions, the ChIP-chip mock summit (midpoint of the highest part of each regions), and the midpoint of each randomly selected sequenced control read.

Motif conservation analysis. PhastCons scores were obtained (as described in the text) for all base pairs for motif occurrences within +/- 150 bp of ChIP-Seq summits and also for those greater than 150 bp but less than 250 bp away from the summits. Number of ChIP-Seq region occurrences for each were CACATG: 396, CACCTG: 74, CACGTG: 63, CAGATG: 173, CAGCTG: 139, CATATG: 105, CA-repeats (3 or more dyads): 610, and GA-repeats (3 or more dyads): 255. A chi squared statistic corresponding to a one-tailed test for a difference between the two distributions was calculated according to the procedure given in Kanji (Kanji 1999 p.83). The two sample sets were first joined and the median for the combined set calculated. The number of PhastCons scores of the background set that were to the left of the combined set median was calculated and designated nl1; the number to the right of the combined median is designated nr1. The two analogous quantities for the ChIP-Seq region motif set were designated nl2 and nr2 with N = nl1+nr1+nl2+nr2. Then the chi squared statistic is calculated as:
N*(| nl1*nr2 – nl2*nr1| - N/2)^2 / ((nl1+nl2)*(nl1+nr1)*(nl2+nr2)*(nr1+nr2))
The x-axis in Fig. 7C represents this test statistic for each motif. Because PhastCons scores are the posterior probability of a given bp to belong to a conserved class of bases, we interpret bp with PhastCons scores > 0.9 as almost certainly conserved. The fraction of bp in ChIP-Seq motifs having PhastCons score > 0.9 is represented as the height of the bars.

Supplemental References

Bailey, A.M. and Posakony, J.W. 1995. Suppressor of hairless directly activates transcription of enhancer of split complex genes in response to Notch receptor activity. *Genes Dev* **9**(21): 2609-2622.

Bailey, T.L., Williams, N., Misleh, C., and Li, W.W. 2006. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* **34**(Web Server issue): W369-373.

Liberman, L.M. and Stathopoulos, A. 2009. Design flexibility in cis-regulatory control of gene expression: synthetic and comparative evidence. *Dev Biol* **327**(2): 578-589.

MacArthur, S., Li, X.Y., Li, J., Brown, J.B., Chu, H.C., Zeng, L., Grondona, B.P., Hechmer, A., Simirenko, L., Keranen, S.V. et al. 2009. Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol* **10**(7): R80.

Markstein, M., Markstein, P., Markstein, V., and Levine, M.S. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proceedings of the National Academy of Sciences* **99**(2): 763.

Markstein, M., Zinzen, R., Markstein, P., Yee, K.P., Erives, A., Stathopoulos, A., and Levine, M. 2004. A regulatory code for neurogenic gene expression in the Drosophila embryo. *Development* **131**(10): 2387-2394.

Ogawa, N., and Biggin, M. D. . 2011. High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro. In *Methods in Mol Biol* (ed. B. Deplanke), p. in press. Humana Press, Clifton, New Jersey.

Pepke, S., Wold, B., and Mortazavi, A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6**(11 Suppl): S22-32.

Roulet, E., Busso, S., Camargo, A.A., Simpson, A.J., Mermod, N., and Bucher, P. 2002. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol* **20**(8): 831-835.

ten Bosch, J.R., Benavides, J.A., and Cline, T.W. 2006. The TAGteam DNA motif controls the timing of Drosophila pre-blastoderm transcription. *Development* **133**(10): 1967.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**(9): R137.

Zinzen, R., Senger, K., Levine, M., and Papatsenko, D. 2006. Computational Models for Neurogenic Gene Expression in the Drosophila Embryo. *Current Biology* **16**(13): 1358-1365.

Chapter 5

Complex interactions between *cis*-regulatory modules in native conformation are critical for *Drosophila snail* expression.

Dunipace L, Ozdemir A, and Stathopoulos A.

RESEARCH ARTICLE 4075

Development 138, 4075-4084 (2011) doi:10.1242/dev.069146

Complex interactions between cis-regulatory modules in native conformation are critical for *Drosophila snail* expression

Leslie Dunipace, Anil Ozdemir and Angelike Stathopoulos*

SUMMARY

It has been shown in several organisms that multiple cis-regulatory modules (CRMs) of a gene locus can be active concurrently to support similar spatiotemporal expression. To understand the functional importance of such seemingly redundant CRMs, we examined two CRMs from the *Drosophila* snail gene locus, which are both active in the ventral region of pre-gastrulation embryos. By performing a deletion series in a ~25 kb DNA rescue construct using BAC recombineering and site-directed transgenesis, we demonstrate that the two CRMs are not redundant. The distal CRM is absolutely required for viability, whereas the proximal CRM is required only under extreme conditions such as high temperature. Consistent with their distinct requirements, the CRMs support distinct expression patterns: the proximal CRM exhibits an expanded expression domain relative to endogenous *snail*, whereas the distal CRM exhibits almost complete overlap with *snail* except at the anterior-most pole. We further show that the distal CRM normally limits the increased expression domain of the proximal CRM and that the proximal CRM serves as a 'damper' for the expression levels driven by the distal CRM. Thus, the two CRMs interact in cis in a non-additive fashion and these interactions may be important for fine-tuning the domains and levels of gene expression.

KEY WORDS: Cis-regulatory modules, Gene expression, *Drosophila melanogaster*, *snail*, Developmental patterning, Repression, Huckebein

INTRODUCTION

A number of cis regulatory modules (CRMs) have recently been identified that support concurrent expression of individual genes in similar spatiotemporal profiles in early *Drosophila* embryos, as well as later in development (e.g. Frankel et al., 2010; Hong et al., 2008; Zeitlinger et al., 2007). For the most part, these secondary CRMs were identified as a result of ChIP-chip and ChIP-seq analyses as regions of occupancy located at a distance from genes of interest, up to 10 kb or more (e.g. Li et al., 2008; Ozdemir et al., 2011; Sandmann et al., 2007; Zeitlinger et al., 2007). These newly identified CRMs have been described as being redundant to previously identified promoter-proximal located CRMs and, most recently, it has been proposed that they function to provide robustness to environmental or genetic perturbation (Frankel et al., 2010; Perry et al., 2010). Moreover, in vertebrate genomes it has been shown that many genes have multiple CRMs active concurrently, and that deletion of one cis-regulatory module can have no observable effect on the gene expression pattern (e.g. Ghiasvand et al., 2011; Xiong et al., 2002). Therefore, identifying why multiple CRMs of similar spatiotemporal expression domains are active simultaneously is a problem of general interest.

Here, we focus on analysis of the *snail* (*sna*) locus in *Drosophila*. *sna* encodes a transcription factor containing Zn-finger DNA-binding domains that predominantly functions to repress the expression of a number of genes from ventral regions of the embryo (e.g. Cowden and Levine, 2002; De Renzis et al., 2006; Ip et al., 1992a). As such, Snail is an important patterning molecule that influences the mesoderm-mesectoderm-neurogenic ectoderm boundary (Kosman et al., 1991; Leptin, 1991). Although a CRM supporting expression similar to *sna* was isolated almost 20 years ago by standard *lacZ* reporter gene constructs from a promoter proximal location, even 6.0 kb of upstream sequence failed to completely represent native *sna* expression, which exhibits very sharp anterior-posterior and lateral boundaries (Ip et al., 1992b). Since then, the predominant view in the field has been that synergy between the Dorsal and Twist transcription factors, which is present in ventral gradients within early embryos, functions to specify the sharp *sna* dorsal boundary (Ip et al., 1992b; Zinzen et al., 2006), and that the sharp posterior boundary is defined by the repressor Huckebein (Reuter and Leptin, 1994). Yet the promoter proximal CRM of *sna* does not exhibit either of these sharp borders, despite the fact that it encompasses the region all the way up to the adjacent upstream gene (Ip et al., 1992b).

In general, it is a common assumption in the field that CRMs located in promoter-proximal locations are required to support gene expression. Thus, although it was noticed that the pattern of the promoter-proximal CRM was expanded relative to endogenous *sna*, the existence of another CRM to serve as a vehicle for repressors was not proposed upon the initial characterization of the reporter gene pattern (Ip et al., 1992b). It is a common finding that CRMs do not always support expression in the exact same domain as the genes they regulate, but in the past this was explained away as a flaw inherent to reporter gene assays. For example, the CRM supporting expression within stripes 3/7 of the *even-skipped* (*eve*) gene does not exhibit equivalent effects in *knirps* mutants as does the endogenous *eve* gene: the expression of the reporter gene expands into the midsection, whereas stripes 3/7 associated with the endogenous *eve* gene retain sharp boundaries (Frasch and Levine, 1987; Small et al., 1996).

Division of Biology, California Institute of Technology, 1200 East California Boulevard, MC114-96, Pasadena, CA 91125, USA.

*Author for correspondence (angelike@caltech.edu)

Accepted 15 July 2011

More recently, however, additional CRMs have been identified sharing similar spatiotemporal profiles to previously characterized CRMs, including one that shares close similarity with the *sna* expression pattern (Ozdemir et al., 2011; Perry et al., 2010). Another recent study presumably labeled this CRM as a 'shadow' enhancer because it is located at a distance from the *snail* gene, whereas the proximally located CRM was defined as the primary acting enhancer (Perry et al., 2010).

To provide insight into the functions of CRMs associated with the *snail* locus in the *Drosophila* early embryo, we undertook a genetic approach towards studying cis-regulatory control using BAC recombineering and site-directed transgenesis to assay the domain and level of expression supported by concurrently functioning CRMs. We focused on the distinction between the proximal and distal *snail* CRMs, which control early embryonic expression, in particular on the patterns and levels of expression supported by each, as well as their abilities to support Snail function.

MATERIALS AND METHODS

Fly stocks

Adh^{nt7}*sna*^{1}*cn*^{1}*vg*^{1}*/CyO*, and *sna*^{18}*/CyO* fly stocks were used (BDSC) after rebalancing with *CyO ftz-lacZ* marked balancer. The proximal 2.2 kb and 6 kb *lacZ* reporter lines and F10 line (hsp83-Toll10B-bcd3′UTR) have been published previously (Huang et al., 1997; Ip et al., 1992b).

Cloning and generation of lacZ constructs

Enhancer sequence for the distal enhancer was amplified from genomic DNA using Sna-Dist 2kb-f (5′-AATTGGTACCACAATTA-GCTGCCGTTTGCAGC-3′) and Sna-Dist 2kb-r (5′-AATTG-GTACCTGTAGCACCCTTGAACTTGTTGTG-3′) and cloned into the *Kpn*I site of the eve$_{promoter}$-*lacZ*-attB vector (Liberman and Stathopoulos, 2009). Site-directed transgenesis system was used to create reporter lines (Bischof et al., 2007). The 86Fb fly stock with attP landing site was injected in house with reporter constructs to generate transgenic lines.

Generation of 25 kb sna rescue constructs

The 25 kb *sna* P[acman] construct was generated using recombineering mediated gap repair performed using SW105 cells as described previously (Venken et al., 2006). The BAC encompassing the *sna* gene (BACR23I04) was obtained from the BacPac Resource Center and the attB-P[acman]-Ap^{R} was modified to contain ~600 bp homology arms to the region of interest. Insertion of GFP just before the stop codon of *sna* was performed using a GFP-frt-kan-frt plasmid and the kan cassette was removed after insertion as described previously (Lee et al., 2001).

Deletion, rearrangement and mutation of the enhancer regions was carried out using the galK system (Warming et al., 2005). All final constructs were isolated and electroporated into EPI300 cells (Epicenter) and the copy number was induced using Fosmid Autoinduction Solution (Epicentre) according to the manufacturers instructions. The constructs were isolated using Nucleobond EF plasmid midi prep kits (Clontech). P[acman] constructs were injected into line 23648 (BDSC) at a concentration of 0.5-1 μg/μl in water using standard techniques. All primers used for gap repair and recombineering are listed in Table S1 in the supplementary material.

Rescue experiment

Lines were created that contained *sna*^{18}*/Cyo ftz-lacZ* and one of the *sna* BAC constructs. Males from these lines were crossed to virgin *Adh*^{nt7}*sna*^{1}*cn*^{1}*vg*^{1}*/CyO ftz-lacZ*. Separate vials were placed at 25°C, 29°C and 18°C. All transgenic flies were counted and the total number of straight wing flies (i.e. *sna* mutants) was compared with the total number of transgenic flies. The final percentage of straight wing flies for each experiment was then divided by 33%, which would be the expected result were the rescue to be perfect.

We note key distinctions between our construct design and that of another recent study of the *snail* locus which used a similar approach (Perry et al., 2010): (1) our transgene functions to rescue a *sna* mutant (i.e. *sna*^{1}*/sna*^{18}) to viability, whereas the other group was limited to assaying early gastrulation defects presumably because a large deficiency background was used; (2) our deletions were guided by our own Twist ChIP-seq data (Ozdemir et al., 2011), effectively guiding definition of the distal CRM as a larger region (~2.0 kb), (3) a spacer sequence (i.e. ampicillin resistance cassette) was not put in place of deletions in our constructs, which allowed us to assay whether native spacing is important; (4) the *sna*-coding sequence, which may possibly influence cis-regulatory mechanism or stability of transcripts, was left intact within our reporter constructs; and (5) the other group did not assay the gastrulation defects associated with the distal CRM delete large transgene but relied on cDNA rescue data conducted previously (Hemavathy et al., 2004).

In situ hybridization

Embryos were fixed and stained following standard protocols. Antisense RNA probes labeled with digoxigenin, biotin or FITC-UTP were used to detect reporter or in vivo gene expression as described previously (Jiang and Levine, 1993; Kosman et al., 2004). Primary antibodies used were: rabbit anti-Eve (provided by M. Frasch, University of Erlangen-Nürnberg, Germany), guinea pig anti-Twist (provided by M. Levine, UC Berkeley, CA, USA), mouse anti-Dorsal (7A4-s from the Hybridoma Bank) and rabbit anti-Histone H3 (Abcam).

Mean intensity quantification

Images of three embryos from each construct were taken using identical parameters. From each embryo, a square of 345 μm^2 was extracted and analyzed for mean intensity using the LSM Image Examiner program (Zeiss). This was repeated three times in each embryo within the *snail* stripe in consistent locations from embryo to embryo. A negative control square of the same size was also analyzed for each embryo. For each measurement within the *snail* stripe, the negative measurement from that embryo was subtracted and then the measurements were averaged and a standard deviation was determined from the nine measurements.

RESULTS

Multiple CRMs in proximity to the *snail* gene support expression in overlapping domains

Previously published Twist-ChIP-seq binding data identified multiple peaks of Twist occupancy to DNA in proximity to the *snail* gene (Fig. 1A) (Ozdemir et al., 2011). By far, the largest peaks were detected ~7 kb upstream of *sna* gene within the intron of another gene, *Tim17b2*. The two proximal Twist occupied regions are covered by the previously studied 2.2 kb and 6 kb enhancer constructs ('proximal CRM') (Ip et al., 1992b). A 2.0 kb DNA fragment from the *Tim17b2* intronic sequence, containing several closely positioned peaks of Twist occupancy, was also assayed in a reporter context ('distal CRM') (Ozdemir et al., 2011).

By analysis of *lacZ* reporter transgenes, we found that both these CRMs (proximal and distal) supported expression in the ventral region of the early embryo in patterns that are spatiotemporally similar but not identical. In contrast to the broadened expression of the proximal CRM fragment (Fig. 1C,F), the distally located CRM fragment supports high-level expression that is refined, sharp and similar to the endogenous *sna* expression pattern (Fig. 1D,G, compare with 1B,E). It should be noted that our tested DNA fragment was defined by Twist ChIP-seq analysis and was larger in size than the one recently tested by another group (i.e. 2 kb versus 1.2 kb) (Perry et al., 2010), a study in which no spatial distinctions between the patterns supported by the proximal and distal CRMs was noted.

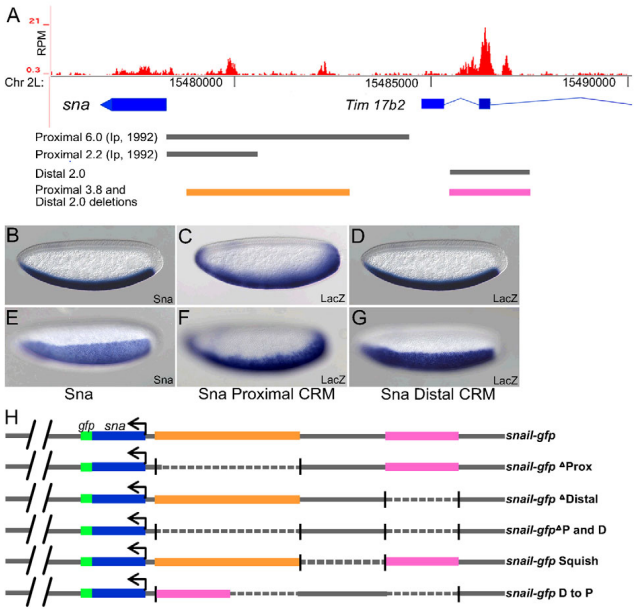


Fig. 1. Distinct regions in the vicinity of the *snail* gene regulate expression in ventral regions of early embryos. (**A**) Twist ChIP-seq defined binding (shown in reads per million, RPM) was identified previously in three domains upstream of *snail*: –1.6, –3.4 and –7 kb (Ozdemir et al., 2011). We created a *lacZ* reporter construct of the ~2 kb distal region in order to encompass the entire region defined by our Twist ChIP-seq analysis, and compared with two *lacZ* reporter constructs assayed previously: proximal 2.2 kb and 6.0 kb constructs (gray lines) (Ip et al., 1992b), regions deleted in the context of a 25 kb rescue construct are shown in orange (proximal) and pink (distal). (**B-G**) In situ hybridization data using riboprobes to detect either *snail* transcript in wild-type embryos (B,E) or *lacZ* transcript in transgenic embryos containing the *snail* 2.2 kb promoter proximal reporter (C,F) or the *snail* distal 2.0 kb reporter (D,G). In this and subsequent figures, embryos are oriented with anterior towards the left. (B-D) Sagittal views; (E-G) ventrolateral surface views. (**H**) A ~25 kb *snail* rescue transgene was modified by insertion of *gfp* as an in-frame fusion to 3′ end of the *snail* gene. Various deletions were created as shown.

Assay of CRM function using larger reporter transgenes in which native context is retained or modified

To analyze how expression of the *sna* gene is controlled in the early embryo, we created a 24.8 kb P[acman] construct encompassing the *sna* gene, as well as flanking DNA sequences using recombineering methods (Fig. 1H) (Venken et al., 2006). We isolated stable transgenic lines using site-directed methods and determined that this DNA sequence can complement the *sna* mutant, suggesting that the cis-regulatory information encoded within this ~25 kb DNA segment is sufficient to support the essential aspects of *sna* expression. To create the reporter construct, we recombineered the *gfp* cDNA sequence into the *sna* locus as an in-frame C-terminal fusion to Snail protein (Fig. 1H, ‘sna-gfp’), allowing us to monitor transgenic expression of *sna-gfp* using a *gfp* riboprobe (see below).

As our goal was to provide insight into cis-regulatory mechanisms regulating *snail* expression, we created five deletion constructs within the 25 kb *sna-gfp* construct using our Twist ChIP-seq data as a guide: (1) a *sna* promoter proximal deletion of 3.8 kb containing two peaks of Twist binding, including most of the 2.2 kb minimal *sna* enhancer identified by Ip et al. (Ip et al., 1992b), but leaving the 500 bp promoter proximal region and including more upstream sequence that we found was also bound by Twist in the early embryo (‘Δ Proximal’); (2) a distal deletion of 2.0 kb, which includes three major peaks of Twist binding, located in the intron of the gene upstream of *sna*, *Tim17b2* (‘Δ Distal’); (3) a double-deletion of both the proximal and distal CRMs (‘Δ P and D’); (4) a deletion of the intervening sequence, present between the proximal and distal CRMs (‘squish’); and (5) a construct in which the distal CRM is moved to the proximal position, in a double-delete background (‘D to P’) (Fig. 1H). 500 bp directly upstream of the *sna*-coding sequence was left unmodified in all cases, with the purpose of leaving the promoter intact.

As both the distal and proximal CRMs supported *sna* expression during early embryogenesis, we investigated whether they function redundantly through analysis of these recombineered reporter transgenes. The proximal CRM deletion (‘Δ Proximal’) supported *gfp* expression that was comparable with *gfp* expression from the full *sna-gfp* rescue construct (Fig. 2B, compare with 2A). Moreover, *gfp* expression similar to that supported by *sna-gfp* was detected in the constructs that moved the distal promoter to a proximal location (‘D to P’) and the construct that deleted the intervening sequence (‘squish’) in the early embryo (data not shown). By contrast, deletion of the distal CRM (‘Δ Distal’) supported weaker expression (Fig. 2C), and the construct that deletes both (‘double delete’) lacked early expression altogether (data not shown). Based on pattern alone, the distal CRM appeared more faithful to the *snail* endogenous expression domain.

Genetic assay of CRM function by *snail* mutant rescue

To determine whether *snail* expression supported by these transgenes was functionally equivalent, we assayed the ability of these transgenes to rescue a *sna* mutant. The wild-type reporter and five modified versions, were introduced into a *sna* mutant background (sna^{1}/sna^{18}) and assayed for their ability to support viability. We found that the native *sna* gene rescued at 91% (Table

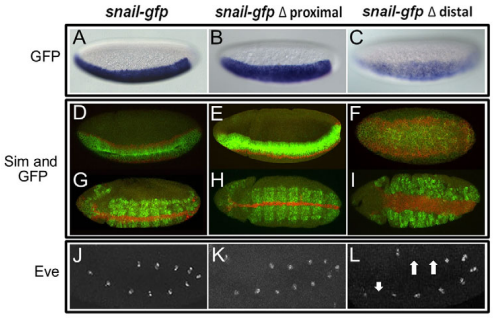


Fig. 2. The distal CRM is required to rescue gastrulation and Eve cell specification defects. (**A-C**) In situ hybridization of cellularized wild-type embryos (stage 5) containing *sna-gfp* construct using a *gfp* riboprobe and alkaline phosphatase staining procedure. *sna-gfp* (A) and *sna-gfp* ΔProximal (B) constructs supported sharp lateral and posterior borders, whereas the *sna-gfp* ΔDistal (C) construct was weaker and exhibited expanded lateral and posterior boundaries. (**D-I**) Fluorescent in situ hybridizations of sna^{1}/sna^{18} mutant embryos using *sim* (red) and *gfp* (green) riboprobes to detect *sna* construct reporter expression and effects on gastrulation through assay of *sim*. *sna* mutant embryos containing either the full-length construct *sna-gfp* (D,G); the proximal delete construct *sna-gfp* ΔProximal (E,H); or the distal delete construct *sna-gfp* ΔDistal (F,I) are shown. (**J-L**) Eve expression in *sna* mutant germ-band elongated embryos containing *sna-gfp* (J), *sna-gfp* ΔProximal (K) or *sna-gfp* ΔDistal (L). Arrows indicate gaps in *eve* expression. (See Fig. S1 in the supplementary material for *sna-gfp* 'D to P' and 'squish images', also see Fig. S2 in the supplementary material for late Eve expression.)

1) but there was significant, but only partial, rescue with the *sna-gfp* fusion constructs (76%) (data not shown). For this reason, we assayed the ability of native *sna* gene constructs, unmodified with *gfp*, to support rescue.

The 25 kb *snail* transgene and the delete proximal CRM constructs rescued the *sna* mutant phenotype; 91% and 82% of expected F1 progeny, respectively, were obtained in rescue crosses (Table 1). By contrast, the distal CRM delete construct completely failed to rescue the *sna* mutant, as did the double delete 'Δ P and D' construct. The 'squish' construct, which removes sequence between the proximal and distal CRMs, also failed to complement the mutant. These results support the conclusion that the distal CRM is required to support viability. In turn, the fact that more than 80% of the expected flies emerged from the *sna* rescue cross with proximal CRM delete transgene suggested that the proximal CRM is not required to support viability.

To further study functional differences between CRMs, we examined the ability of our constructs to support viability at various temperatures: 25°C, 29°C and 18°C. The proximal CRM delete construct showed decreased viability at higher temperature, with 36% viability supported at 29°C when compared with 82% at 25°C; yet at 18°C, we found the rescue was also high at 94% (Table 1). However, we found that the distal CRM delete construct did not rescue at any temperature tested: 0% viability at 18°C, 25°C, and 29°C; further evidence that the distal CRM is the primary CRM responsible for supporting *sna* expression.

Deletion of the distal CRM, specifically, has consequences on gastrulation

Next, we examined whether these CRMs have similar or different roles during gastrulation. The constructs containing the distal CRM rescued the gastrulation defects of *sna* mutants [i.e. 'Δ Proximal' (Fig. 2E,H) and 'squish' and 'D to P' (see Fig. S1 in the supplementary material), compare with full length *sna-gfp* (Fig. 2D,G)]. By contrast, constructs without the distal CRM exhibited gastrulation defects (i.e. 'Δ Distal', Fig. 2F,I). In the absence of the distal CRM, not only was *single-minded* (*sim*) expression aberrant, with expansion into a broad domain compared with the single line of cells found in wild-type embryos, but invagination was non-uniform and presumably contributed to unequal mesoderm spreading (Fig. 2F,I). As *sim* is directly repressed by the Snail transcription factor in gastrulating embryos (Kasai et al., 1992), these results indicated that the level of *snail* expression in the *sna* mutant background supported by *sna-gfp* Δ Distal is insufficient to fully support function at this stage of development, resulting in an expansion of the *sim* domain.

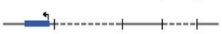
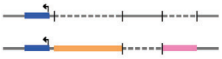
As an assay for possible later phenotypes, we examined expression of *even-skipped* (*eve*). *eve* encodes a homeodomain transcription factor necessary for dorsal mesoderm lineage specification (Frasch et al., 1987), and its lateral expression in 11 clusters of cells on either side of the embryo at stage 11 can be used as an indicator for proper mesoderm spreading. In rescue experiments in which the distal CRM was absent, *eve* expression was aberrant as gaps in expression were detected in all of the embryos examined (Fig. 2L, arrows). By contrast, constructs that removed the proximal CRM, leaving the distal CRM intact, exhibited normal gastrulation (invagination and *sim* expression, Fig. 2E,H), as well as normal mesoderm spreading and specification even at later stages of embryogenesis (Eve expression; Fig. 2K). Even when the temperature was raised to 29°C, no obvious mesoderm specification defects in the trunk of the embryos were observed in the absence of the proximal CRM (see Fig. S2 in the supplemental material). Our data for rescue of the sna^{1}/sna^{18} background demonstrated that the distal CRM is required to support gastrulation, but that the proximal CRM is not required or supports a minor role (such as supporting expression at the anterior, see below).

The proximal CRM deletion of 3.8 kb removes multiple tissue-specific enhancers, a minimum of three: one module from 1.2 kb to 2 kb supports expression in ventral regions of the early embryo (e.g. Fig. 1C) and two other modules, one from 0.4-0.9 kb and another from 2.2-2.8 kb, support expression in the peripheral nervous system (PNS) and central nervous system (CNS), respectively, at later stages of embryogenesis (Ip et al., 1994; Ip et al., 1992b). We observed changes in the PNS and CNS expression in constructs that delete the proximal CRM, but no effect on expression in these domains was observed in the constructs that delete the distal CRM (see below).

Multiple CRMs support *sna* expression in germ-band elongated embryos and are organized on the chromosome in a manner that potentially minimizes dominant effects of repressors

In the course of our *sna* rescue experiments, we found that a construct removing the intervening sequence between distal and proximal CRMs was not able to complement the mutant (Table 1,

Table 1. The distal enhancer is required for viability at all temperatures, whereas the proximal enhancer is required conditionally at high temperatures

		Percentage rescue		
	Transgene	25°C	18°C	29°C
	Sna rescue construct	91% (*n*=170)	100% (*n*=52)	100% (*n*=23)
	Sna Δ proximal CRM	82% (*n*=51)	94% (*n*=29)	36% (*n*=34)
	Sna Δ distal CRM	0% (*n*=44)	0% (*n*=18)	0% (*n*=22)
	Sna Δ proximal and distal	0% (*n*=47)		
	Sna squish	0% (*n*=95)		

Schematics of each of the constructs are shown on the left. Percentage rescue indicates the number of sna^{1}/sna^{18} flies counted out of the total number of flies present, then divided by what would be considered a complete rescue (i.e. 33% of total flies). *n* is the total number of flies counted. Because the 'Δ proximal and distal' and 'squish' constructs did not rescue at 25°C, they were not further analyzed at the other temperatures.

'*sna-gfp* squish'). We hypothesized that either this sequence supports another function required for viability or it influences the ability of the distal enhancer to function. To test the first possibility, we examined expression of *sna* in slightly older embryos, ones that were undergoing germ-band elongation. Previous studies have documented *sna* expression at this stage within the ectoderm and in malphigian tubule (MT) precursor cells (Alberga et al., 1991; Ip et al., 1994). From analysis of germ-band elongated embryos (stage 9), we observed that *sna* was also expressed at this stage in the posterior midgut (PMG) and in the head (possibly marking either anterior midgut and/or head mesoderm) (Fig. 3A) (Alberga et al., 1991; de Velasco et al., 2006).

The patterns of reporter expression supported by each *sna-gfp* transgene were analyzed (Fig. 3B). When the proximal CRM region was deleted, we found that a subset of expression in the ectoderm was lost (i.e. pattern 'Ect1') (Ip et al., 1994). Yet upon loss of the 3.8 kb proximal CRM, expression in the neurogenic ectoderm was retained in stripes within the trunk but was absent in the midsection domain of the embryo (i.e. pattern 'Ect2'), suggesting that other sequences also impact ectodermal expression. We deduced that the CRM responsible for supporting expression in the Ect2 pattern is most probably present in the DNA sequence of our rescue construct downstream of *sna* (~14 kb), because none of the modified constructs we tested ever affected expression of the reporter in this domain. Next, we found that expression within the MT precursor cells was completely lost when the distal CRM was deleted (Fig. 3B, delete distal: 'Δ distal' and double delete: 'Δ P and D') and that the pattern was retained as long as the distal CRM was present, even if located in a different location. When the distal CRM was moved to the proximal position ('D to P'), there was an overall diminishment of expression in all domains but the MT precursor cell expression was retained. These results suggested that the 2.0 kb DNA associated with the distal CRM supports expression in the MT precursor cells in addition to its function in supporting early *sna* expression in ventral regions of the embryo. Consistent with this view, when the distal CRM *lacZ* construct was examined, expression in MT precursor cells within embryos at stage 9 was also observed (data not shown).

Last, the 'squish' construct was the only construct found to cause loss of expression in the head and PMG, suggesting that this intervening sequence contains CRMs that support these *sna* expression domains. Loss of expression in these domains may be responsible for the inability of this construct to rescue the mutant. The 'squish' construct also resulted in a partial to complete loss of expression within the Ect1 region (Fig. 3B, gray box). Although it is possible that deletion of the intervening sequence from –4.3 to –7.2 kb, which was removed by the 'squish', could influence neuronal expression; this is unlikely as full *sna* expression within the CNS and PNS is observed with a transgene that includes only the most proximal 2.8 kb (Ip et al., 1994).

We hypothesized that by moving the two CRMs into closer proximity by deleting the intervening DNA ('squish'), repressors acting within the distal CRM may function to repress expression in the ectoderm normally supported by the proximal CRM. This idea, together with the fact that the distal CRM exhibited spatially refined expression relative to the proximal CRM in the early embryo (e.g. Fig. 1B-G), led us to investigate whether repressor(s) that act to limit *snail* expression function through the distal CRM.

Repressors predominantly function through the distal CRM to regulate the posterior and dorsal boundaries of the *sna* expression domain within the early embryo

It has previously been shown that the Huckebein (Hkb) transcription factor, which is expressed at both the anterior and posterior poles, functions as a repressor to define the posterior boundary of *sna* expression (Goldstein et al., 1999; Reuter and Leptin, 1994). In *hkb* mutants, posterior *sna* expression is expanded into the pole and anterior expression is expanded beyond the tip and into the dorsal region of the embryo. Upon examination of the *sna-gfp* construct in which the proximal CRM was deleted, we found that *gfp* expression was excluded from the posterior *hkb* expression domain, similar to endogenous *sna* expression (Fig. 4B, compare with 4A). This result suggested that Hkb can function to repress the *sna* posterior boundary, even when the proximal CRM is removed. By contrast, *gfp* expression was expanded into the posterior end of the embryo upon deletion of the distal CRM (Fig. 4C, compare with 4A).

sna and *hkb* expression domains overlap at anterior regions of the embryo. Upon closer analysis of the *sna-gfp* proximal delete construct, we found that the *gfp* expression domain recedes relative to *sna*, such that the boundary of expression was more ventrally located and sharper relative to wild type (Fig. 4E). A similar effect on *sna* expression has been observed previously in *bicoid* mutants (Reuter and Leptin, 1994). However, in comparison with the expression domain supported by the *sna-gfp* distal CRM delete, we found that the *sna* expression domain was expanded more dorsally at the anterior of the embryo than normal (Fig. 4F), similar to that seen in *hkb* mutants (Reuter and Leptin, 1994). Collectively, these results suggest the proximal CRM supports Bicoid-mediated activation at the anterior of the embryo and that the distal CRM supports Hkb-mediated repression at both embryonic poles.

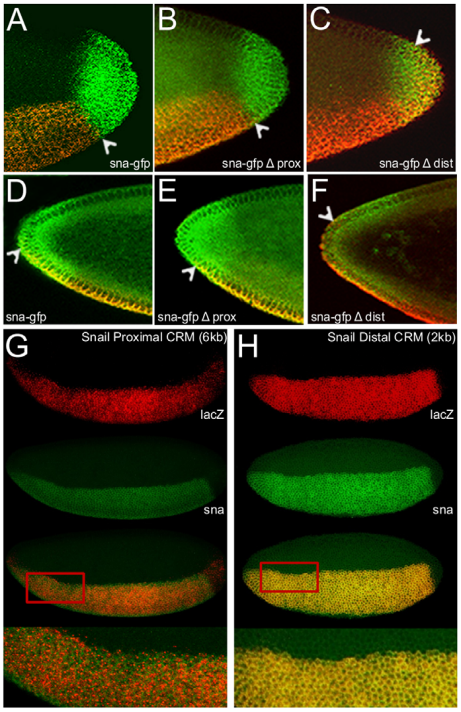


Fig. 4. Repressors function predominantly through the distal CRM, whereas expansion toward the anterior pole requires the proximal CRM. (**A-F**) Fluorescent in situ hybridization of wild-type embryos (stage 5) containing either the *sna-gfp* (A,D), *sna-gfp* Δ Proximal (B,E) or *sna-gfp* Δ Distal (C,F) constructs using riboprobes to detect *gfp* (red) and *hkb* (green) transcripts. Magnified images of the poles of stage 5 embryos showing the posterior (A-C) and anterior (D-F) variation in *sna-gfp* expression (red) with respect to the domain of *hkb* expression (green). The posterior images are projections, whereas the anterior images represent a single scan. Extent of *gfp* expression supported at the poles is marked by arrowheads in each case. (**G**,**H**) Ventrolateral views of in situ hybridization recognizing *lacZ* (red), driven by either the proximal CRM (G) or the distal CRM (H), and *snail* (green). The red rectangle in each indicates the area magnified in the bottom image.

CRMs. We hypothesized that repressors associated with the distal CRM might also work to define the expression supported by the proximal CRM output. This would explain why the endogenous *snail* expression domain was absent from the posterior pole and also why its lateral boundary was sharp. However, it was also possible that the level of expression supported by each CRM was so different that when both were present, the pattern supported by the distal CRM effectively masked that supported by the proximal CRM. To distinguish between these possibilities, we examined embryos containing various combinations of the proximal delete and/or the distal delete CRM reporters, in either cis or trans conformation, and analyzed the gene expression outputs supported by each combination in terms of spatial domain (Fig. 5) and level of expression (Fig. 6).

At two copies, the proximal CRM delete construct supported refined expression (repressed at the posterior and laterally), whereas the distal CRM delete construct supported expanded expression (extending at the poles and laterally) compared with an unmodified reporter construct (Fig. 5A,B, compare with 5D), similar to expression supported by one copy of the transgenes. However, when reporter expression was assayed in an embryo containing one copy of the proximal CRM delete and one copy of the distal CRM delete transgenes, the pattern supported exhibited an expanded expression domain, most apparent at the posterior pole. This result suggested that the expression supported by the proximal CRM is not simply too weak to be observed in the presence of the expression supported by the distal CRM, but that instead repressors associated with the distal CRM normally function to refine expression at the poles and in lateral regions supported by the proximal CRM. Furthermore, these data demonstrate that repressors associated with the distal CRM cannot function in trans, but instead require a cis conformation relative to position of the proximal CRM in order to have an effect. Our results suggest that the normal pattern is a non-additive reflection of the domains of expression supported by each CRM (see Discussion).

Besides differences in domain of expression, we noticed that these constructs supported differences in levels of expression (Fig. 6). When imaged at a power and gain in which all of the constructs examined were not over-exposed, the mean intensity supported by the *sna-gfp* and *sna-gfp Δ distal* constructs were comparable, but in comparison the expression levels supported by the *sna-gfp Δ proximal* construct were considerably higher (~3-4 fold). Therefore, in the absence of the proximal CRM, the expression levels increased. At higher gain, however, it was observed that the *sna-gfp* expression was at least twofold higher than that of the *sna-gfp Δ distal* (data not shown). Thus, alternately, in the absence of the distal CRM, the expression levels decreased. In addition, the *sna-gfp squish* construct also supported increased levels of expression relative to the *sna-gfp* construct (approximately twofold). Collectively, these results suggest that normal levels and patterns of *snail* gene expression require input from both the proximal and distal CRMs, and that effective regulation of expression levels requires proper organization of these CRMs upon the chromosome.

DISCUSSION

In this study, we provide evidence that early *snail* expression is regulated by two concurrently acting CRMs that support gene expression patterns that are spatially and functionally different. The distally located CRM is necessary to support gastrulation as well as viability of *snail* mutants, whereas the proximal CRM is dispensable for viability except at high temperature. Furthermore, our data show these CRMs support distinct expression patterns. Although they probably share many transcription factors, the distal CRM alone is responsive to the repressor Huckebein and the unknown laterally acting ‘repressor X’, whereas the proximal CRM alone responds to an anterior activator.

Our data suggest that the proximal CRM functions as a ‘damper’ to reduce the high levels of expression normally supported by the distal CRM. Multiple CRMs associated with a single gene may support spatiotemporally similar expression patterns, but the mean levels of gene expression supported by each can be very different.

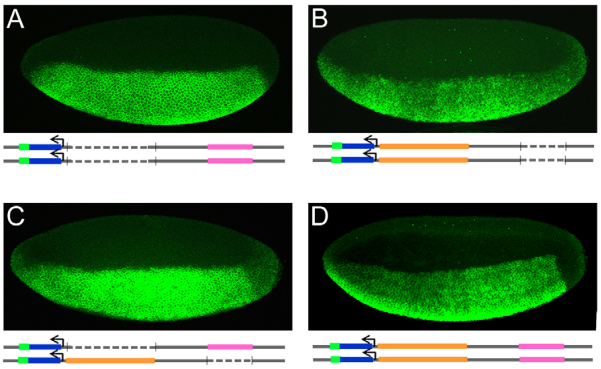


Fig. 5. The proximal and distal enhancers function in a non-additive manner when organized in cis conformation but not in trans. (**A-D**) Fluorescent in situ hybridization using a *gfp* riboprobe of stage 5 embryos expressing one of the following constructs: homozygous 'snail-gfp Δ proximal' (A), homozygous 'snail-gfp Δ distal' (B), heterozygous 'snail-gfp Δ proximal'/'snail-gfp Δ distal' (C) or homozygous 'snail-gfp' (D). All images were captured under the same confocal settings but the brightness and/or contrast was modulated to support visual comparison of the domains of expression supported by these four transgenes.

In the case of the *snail* locus, our data show that the distal and proximal CRMs drive high or low levels of expression, respectively, within a similar domain in ventral regions of the embryo. Our results supports a model in which these two CRMs provide dual-control of expression levels, high versus low, to provide flexibility in terms of levels of *snail* expression (Fig. 6F). The requirement for the proximal CRM at high temperatures could indicate a need to more closely regulate the expression levels of *snail* in stressful environments. Such flexibility is probably advantageous and may explain why two CRMs that support similar expression patterns may be evolutionarily constrained.

Both the proximal and distal CRMs support expression not only during gastrulation in ventral regions of the embryo but in other domains at later stages of development. The distal CRM also supports expression within malphigian tubule precursors (Fig. 3), and, as was previously shown, the proximal CRM supports expression later within neuroblasts (Ip et al., 1992b). Therefore, these elements can be reused during the course of development, and may be evolutionarily retained for reasons beyond a role in canalization.

CRMs associated with the *snail* locus function in a non-additive manner to support expression

Our results show that transcription factors associated with the distal CRM can dominantly affect the other proximally located CRM to support expression of *sna* that is refined and excluded from the posterior pole. Our data support the view that non-autonomous CRM function is responsible for the resulting pattern which is effectively non-additive, i.e. it is not simply the summed equivalent of the domains of expression supported by the two CRMs. Non-autonomous CRM function may be advantageous, providing additional flexibility by allowing individual and combined activities of CRMs based on circumstances, to support canalization. It has been demonstrated that non-additive CRM interactions also play a role defining the expression domain of another *Drosophila* early patterning gene, *sloppy-paired 1* (Prazak et al., 2010). Our data support the view that this is a more common cis-regulatory mechanism than currently appreciated. For example, even in case of the *even-skipped* gene locus that has received considerably focus, questions remain about why particular CRM behaviors are not equivalent to the behaviors of the *eve* gene itself. The expansion of a *eve* stripe 3/7 reporter gene in *knirps* mutants (Small et al., 1996), but not the *eve* gene itself (Frasch and Levine, 1987), suggests that another repressor is required to drive proper *eve* stripe 3/7 expression and that this activity is supported through another DNA fragment. We propose that another CRM associated with the *eve* locus may aid in definition of *eve* stripes 3/7 by serving as a vehicle for additional repressors(s), similar in mechanism to regulation of *snail* gene expression shown here in this study.

CRMs are organized along the DNA to support effective transcription

This study also supports the view that CRMs are organized in the context of the gene locus to support proper patterning and to minimize cross-repressive interactions (see also Cai et al., 1996; Small et al., 1993). We believe that the loss of Ect1 expression that we see in the 'squish' construct is the result of dominant repression, owing to the fact that the distal enhancer is moved in proximity to the proximal enhancer (see Fig. 3B). This would suggest that the native context of CRMs within a locus can limit interactions between elements, and may go towards explaining why enhancers in diverged species/animals tend to be found in the same general location (Cande et al., 2009; Hare et al., 2008). Similarly, the dampening of all *snail* expression patterns we observe in the 'D to P' construct may be due to the repressive activity of the distal CRM being moved near the promoter.

Placing binding sites for repressors near the promoter potentially limits the range of activity of a gene. Many genes involved in early development, such as *snail*, take on different roles later in development and are subject to different molecular inputs during the life of the animal. Like *snail*, the *intermediate neuroblasts defective* (*ind*) gene also has a distally located enhancer and another that is located in the proximal position. Similar to what we see at the *snail* locus, the distal CRM has documented repression associated with it, whereas the proximally located element functions through positive autoregulatory feedback (Stathopoulos and Levine, 2005; Von Ohlen et al., 2007). We suggest that keeping repressors located at a distance from the promoter supports flexibility in reiterative reactivation of genes throughout the course of development. However, in addition to buffering repressive crosstalk through distance, we propose that linking repression function to the presence of an activator (i.e. between CRMs concurrently active in the same cells) may also serve as an alternate mechanism to moderate non-autonomous CRM interactions; other

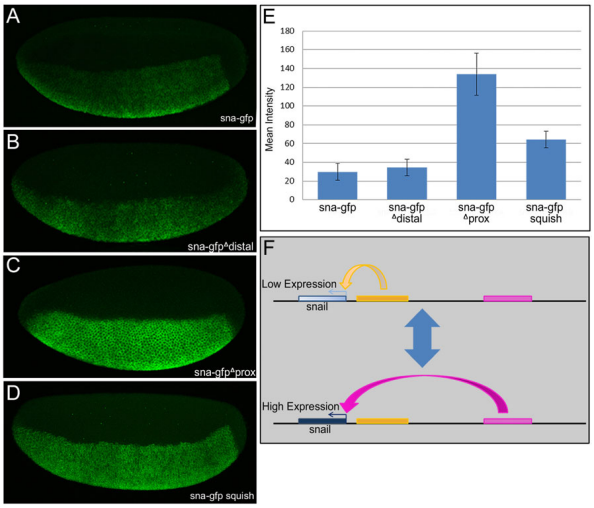


Fig. 6. The proximal and distal CRMs influence levels of *snail* expression. (**A-D**) Fluorescent in situ hybridization using a *gfp* riboprobe to recognize expression from the *gfp* reporter constructs. All images were captured under the same settings and there has been no manipulation of the brightness or contrast to allow for visual comparison of the expression levels supported by these four transgenes. (**E**) The mean intensity of expression in the *snail* stripe for all four constructs was quantified, showing that deletion of the proximal CRM leads to a greater than fourfold difference in expression levels. Data are mean±s.e.m. (**F**) Schematic of the interplay between the proximal and distal CRMs and the *snail* promoter. The proximal CRM drives low level expression, whereas the distal CRM drives high level expression; there is most probably a trade off between the two CRMs, effectively lowering the level of expression that is seen in the full *snail* construct to a level many times lower than that supported by the distal CRM alone.

studies in the past have suggested that repressors may require activators to bind DNA (i.e. 'hot chromatin' model) (see Nibu et al., 2001).

Our data show that expression of the *Drosophila snail* gene in embryos is established through integrated activity of multiple CRMs that function concurrently and, in part, through non-additive interactions. Non-additive activity of CRMs, through sharing of repressors for example, is likely more commonplace than currently appreciated. It is possible that concurrently acting CRMs function coordinately to regulate spatial domain and levels of expression in general, and may provide one explanation why genes in *Drosophila* and other animals often have multiple CRMs that support similar spatiotemporal patterns of expression.

Acknowledgements

We are grateful to Eric Davidson (Caltech) and to the Stathopoulos lab for helpful discussions. In addition, we thank Sagar Damle (Caltech) for advice on BAC recombineering. This work was funded through NIGMS R01 grant GM077668 (A.S.) and corresponding ARRA supplement from the NIH. Deposited in PMC for release after 12 months.

Competing interests statement

The authors declare no competing financial interests.

Supplementary material

Supplementary material for this article is available at
http://dev.biologists.org/lookup/suppl/doi:10.1242/dev.069146/-/DC1

References

Alberga, A., Boulay, J. L., Kempe, E., Dennefeld, C. and Haenlin, M. (1991). The snail gene required for mesoderm formation in *Drosophila* is expressed dynamically in derivatives of all three germ layers. *Development* **111**, 983-992.

Bischof, J., Maeda, R. K., Hediger, M., Karch, F. and Basler, K. (2007). An optimized transgenesis system for *Drosophila* using germ-line-specific {varphi}C31 integrases. *Proc. Natl. Acad. Sci. USA* **104**, 3312.

Cai, H. N., Arnosti, D. N. and Levine, M. (1996). Long-range repression in the *Drosophila* embryo. *Proc. Natl. Acad. Sci. USA* **93**, 9309-9314.

Cande, J. D., Chopra, V. S. and Levine, M. (2009). Evolving enhancer-promoter interactions within the tinman complex of the flour beetle, Tribolium castaneum. *Development* **136**, 3153-3160.

Cowden, J. and Levine, M. (2002). The Snail repressor positions Notch signaling in the *Drosophila* embryo. *Development* **129**, 1785-1793.

De Renzis, S., Yu, J., Zinzen, R. and Wieschaus, E. (2006). Dorsal-ventral pattern of Delta trafficking is established by a Snail-Tom-Neuralized pathway. *Dev. Cell* **10**, 257-264.

de Velasco, B., Mandal, L., Mkrtchyan, M. and Hartenstein, V. (2006). Subdivision and developmental fate of the head mesoderm in *Drosophila melanogaster*. *Dev. Genes Evol.* **216**, 39-51.

Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F. and Stern, D. L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**, 490-493.

Frasch, M. and Levine, M. (1987). Complementary patterns of even-skipped and fushi tarazu expression involve their differential regulation by a common set of segmentation genes in *Drosophila*. *Genes Dev.* **1**, 981-995.

Frasch, M., Hoey, T., Rushlow, C., Doyle, H. and Levine, M. (1987). Characterization and localization of the even-skipped protein of *Drosophila*. *EMBO J.* **6**, 749-759.

Ghiasvand, N. M., Rudolph, D. D., Mashayekhi, M., Brzezinski, J. A. t., Goldman, D. and Glaser, T. (2011). Deletion of a remote enhancer near ATOH7 disrupts retinal neurogenesis, causing NCRNA disease. *Nat. Neurosci.* **14**, 578-586.

Goldstein, R. E., Jimenez, G., Cook, O., Gur, D. and Paroush, Z. (1999). Huckebein repressor activity in *Drosophila* terminal patterning is mediated by Groucho. *Development* **126**, 3747-3755.

Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R. and Eisen, M. B. (2008). Sepsid even-skipped enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. *PLoS Genet.* **4**, e1000106.

Hemavathy, K., Hu, X., Ashraf, S. I., Small, S. J. and Ip, Y. T. (2004). The repressor function of snail is required for *Drosophila* gastrulation and is not replaceable by Escargot or Worniu. *Dev. Biol.* **269**, 411-420.

Hong, J.-W., Hendrix, D. A. and Levine, M. S. (2008). Shadow enhancers as a source of evolutionary novelty. *Science* **321**, 1314.

Huang, A. M., Rusch, J. and Levine, M. (1997). An anteroposterior Dorsal gradient in the *Drosophila* embryo. *Genes Dev.* **11**, 1963-1973.

Ip, Y. T., Park, R. E., Kosman, D., Bier, E. and Levine, M. (1992a). The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive neuroectoderm of the *Drosophila* embryo. *Genes Dev.* **6**, 1728-1739.

Ip, Y. T., Park, R. E., Kosman, D., Yazdanbakhsh, K. and Levine, M. (1992b). dorsal-twist interactions establish snail expression in the presumptive mesoderm of the *Drosophila* embryo. *Genes Dev.* **6**, 1518-1530.

Ip, Y. T., Levine, M. and Bier, E. (1994). Neurogenic expression of snail is controlled by separable CNS and PNS promoter elements. *Development* **120**, 199-207.

Jiang, J. and Levine, M. (1993). Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen. *Cell* **72**, 741-752.
Jiang, J., Kosman, D., Ip, Y. T. and Levine, M. (1991). The dorsal morphogen gradient regulates the mesoderm determinant twist in early *Drosophila* embryos. *Genes Dev.* **5**, 1881.
Kasai, Y., Nambu, J. R., Lieberman, P. M. and Crews, S. T. (1992). Dorsal-ventral patterning in *Drosphila*: DNA binding of *snail* protein to the single-minded gene. *Proc. Natl. Acad. Sci. USA* **89**, 3414-3418.
Kosman, D., Ip, Y. T., Levine, M. and Arora, K. (1991). Establishment of the mesoderm-neuroectoderm boundary in the *Drosophila* embryo. *Science* **254**, 118-122.
Kosman, D., Mizutani, C. M., Lemons, D., Cox, W. G., McGinnis, W. and Bier, E. (2004). Multiplex detection of RNA expression in *Drosophila* embryos. *Science* **305**, 846.
Lee, E. C., Yu, D., Martinez de Velasco, J., Tessarollo, L., Swing, D. A., Court, D. L., Jenkins, N. A. and Copeland, N. G. (2001). A highly efficient Escherichia coli-based chromosome engineering system adapted for recombinogenic targeting and subcloning of BAC DNA. *Genomics* **73**, 56-65.
Leptin, M. (1991). twist and snail as positive and negative regulators during *Drosophila* mesoderm development. *Genes Dev.* **5**, 1568-1576.
Li, X. Y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D. A., Iyer, V. N., Hechmer, A., Simirenko, L., Stapleton, M., Luengo Hendriks, C. L. et al. (2008). Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol.* **6**, e27.
Liberman, L. M. and Stathopoulos, A. (2009). Design flexibility in cis-regulatory control of gene expression: synthetic and comparative evidence. *Dev. Biol.* **327**, 578-589.
Nibu, Y., Zhang, H. and Levine, M. (2001). Local action of long-range repressors in the *Drosophila* embryo. *EMBO J.* **20**, 2246-2253.
Ozdemir, A., Fisher, K., Pepke, S., Samanta, M., Dunipace, L., McCue, K., Zeng, L., Ogawa, N., Wold, B. and Stathopoulos, A. (2011). High resolution mapping of Twist to DNA in *Drosophila* embryos: efficient functional analysis and evolutionary conservation. *Genome Res.* **21**, 566-577.
Perry, M. W., Boettiger, A. N., Bothma, J. P. and Levine, M. (2010). Shadow enhancers foster robustness of *Drosophila* gastrulation. *Curr. Biol.* **20**, 1562-1567.
Prazak, L., Fujioka, M. and Gergen, J. P. (2010). Non-additive interactions involving two distinct elements mediate sloppy-paired regulation by pair-rule transcription factors. *Dev. Biol.* **344**, 1048-1059.
Reuter, R. and Leptin, M. (1994). Interacting functions of snail, twist and huckebein during the early development of germ layers in *Drosophila*. *Development* **120**, 1137-1150.
Sandmann, T., Girardot, C., Brehme, M., Tongprasit, W., Stolc, V. and Furlong, E. E. (2007). A core transcriptional network for early mesoderm development in *Drosophila melanogaster*. *Genes Dev.* **21**, 436-449.
Small, S., Arnosti, D. N. and Levine, M. (1993). Spacing ensures autonomous expression of different stripe enhancers in the even-skipped promoter. *Development* **119**, 762-772.
Small, S., Blair, A. and Levine, M. (1996). Regulation of two pair-rule stripes by a single enhancer in the *Drosophila* embryo. *Dev. Biol.* **175**, 314-324.
Stathopoulos, A. and Levine, M. (2005). Localized repressors delineate the neurogenic ectoderm in the early *Drosophila* embryo. *Dev. Biol.* **280**, 482-493.
Venken, K. J., He, Y., Hoskins, R. A. and Bellen, H. J. (2006). P[acman]: a BAC transgenic platform for targeted insertion of large DNA fragments in *D. melanogaster*. *Science* **314**, 1747-1751.
Von Ohlen, T. L., Harvey, C. and Panda, M. (2007). Identification of an upstream regulatory element reveals a novel requirement for Ind activity in maintaining ind expression. *Mech. Dev.* **124**, 230-236.
Warming, S., Costantino, N., Court, D. L., Jenkins, N. A. and Copeland, N. G. (2005). Simple and highly efficient BAC recombineering using galK selection. *Nucleic Acids Res.* **33**, e36.
Xiong, N., Kang, C. and Raulet, D. H. (2002). Redundant and unique roles of two enhancer elements in the TCRgamma locus in gene regulation and gammadelta T cell development. *Immunity* **16**, 453-463.
Zeitlinger, J., Zinzen, R. P., Stark, A., Kellis, M., Zhang, H., Young, R. A. and Levine, M. (2007). Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the *Drosophila* embryo. *Genes Dev.* **21**, 385-390.
Zinzen, R. P., Senger, K., Levine, M. and Papatsenko, D. (2006). Computational models for neurogenic gene expression in the *Drosophila* embryo. *Curr. Biol.* **16**, 1358-1365.

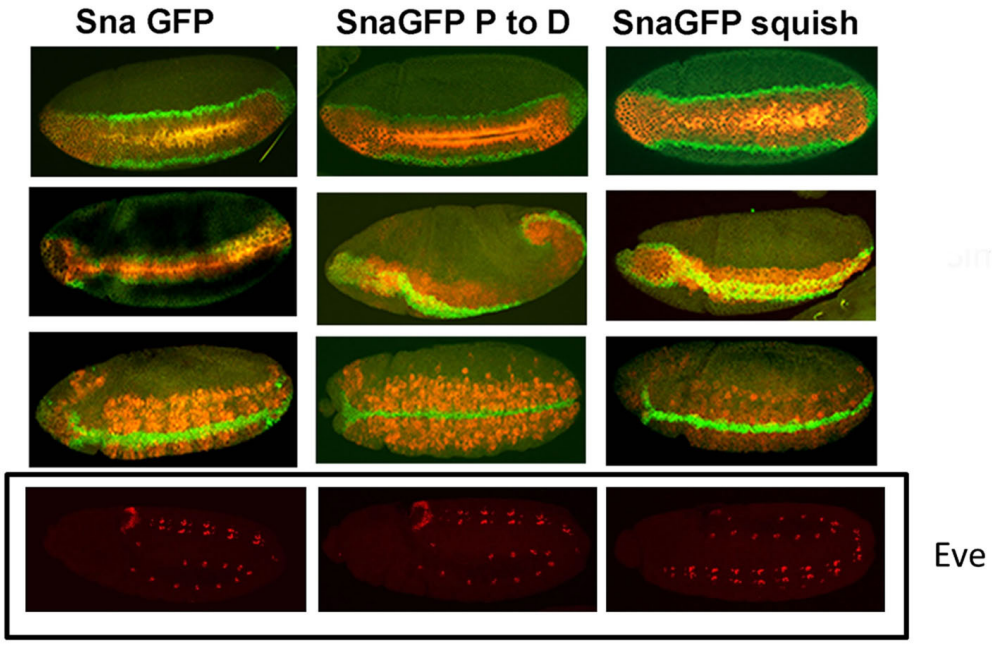


Fig. S1. All BAC constructs containing the distal enhancer are able to rescue sna mutant gastrulation defects. (A-F) Florescent in situ hybridizations with *sim* (green) and *gfp* (red) on sna^{1}/sna^{IIGo5} embryos expressing the indicated BAC show that *sim* expression during and after gastrulation is essentially normal in both the squish and the D to P constructs. Antibody staining using an anti-Eve antibody demonstrates that Eve expression (red, G-I) is also normal, showing 11 clusters of Eve positive cells on the lateral side of the embryo, indicating that mesoderm spreading was normal in these embryos.

Fig. S2. Even at high temperatures the distal enhancer is sufficient to rescue gastrulation in sna mutants. In situ hybridization for *gfp* (green) and antibody staining to detect Eve (red) show that even at 29°C the phenotype of the delete proximal and delete distal CRM constructs are very different. The delete proximal construct appears largely normal, with the mesodermal Eve cells present at the dorsal side of the embryo. The delete distal construct exhibits major defects, with only 4 clusters of Eve+ cells at the dorsal side of the embryo.

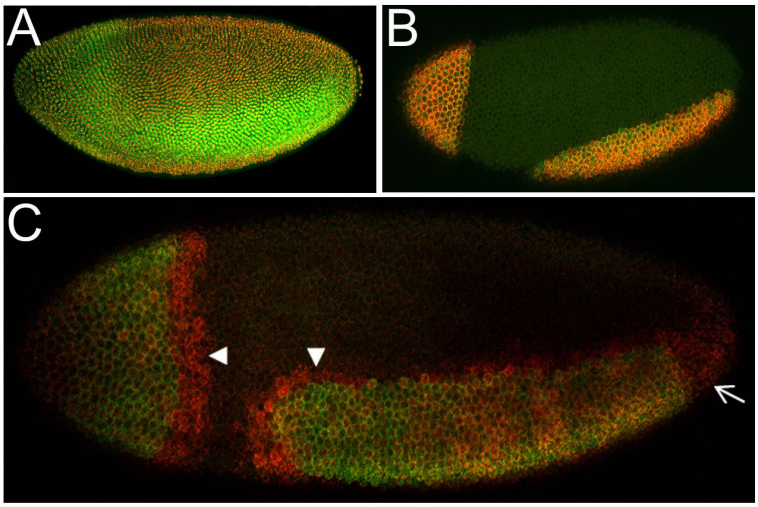


Fig. S3. Repressors function predominantly through the distal CRM, whereas expansion towards the anterior pole requires the proximal CRM. Embryos containing two intersecting dorsal-ventral patterning axes supported by expression of activated Toll receptor at the anterior of wild-type embryos accomplished through a transgene, F10 (Huang et al., 1997). (A) F10-expressing embryos were stained using an anti-Dorsal (green) and anti-Histone H3 (red) antibodies to identify dorsal-ventral patterning axes and nuclei (for image contrast), respectively. (B,C) In situ hybridization using riboprobes to *lacZ* (red) and *sna* (green) of embryos in which either the distal 2.0 kb CRM (B) or the proximal 2.2 kb CRM (C) *lacZ* reporter constructs were introduced into a background containing the F10 transgene. In B, lacZ and *sna* expression completely overlap, showing that the distal CRM is responsive to repressor X. In C, lacZ expression was found to extend well beyond the expression domain of sna expression into the posterior pole (large arrow) as well as into the domain of repressor X function (arrowheads).

Histones are subject to post-translational modifications that have been linked to a variety of cellular processes. Histone modifications that occur in the N-terminal tails have been thought to mediate binding of non-histone proteins and protein complexes to chromatin. However, some modifications, such as acetylation and phosphorylation, can alter the charge of histone tails and therefore, have the potential to influence chromatin through electrostatic mechanisms. Here we show that lysine 56 within the core domain of histone H3 is acetylated. Our data suggest that H3-K56Ac might play a role in DNA damage repair as the yeast cells carrying the K56R mutant allele showed sensitivity to a variety of DNA damaging agents. We generated an antibody specific for this modification to identify the acetylase responsible for acetylation of H3-K56. Our screen using deletion strains for 16 known yeast protein acetylases showed that H3 lysine 56 acetylation levels were unaffected by the loss of any single enzyme suggesting either that an as yet unidentified HAT exists or that multiple HATs can acetylate H3-K56. Using antibody staining, we showed that K56 acetylation levels increase during S-phase whereas there is substantially less K56 acetylation in G1 cells, suggesting that this modification might be incorporated into chromatin during histone deposition.

Gene expression is regulated at the transcriptional level by *cis*-regulatory modules (CRMs) that are bound by sequence-specific transcription factors (TFs). Recent studies have identified *in vivo* binding profiles of several TFs across the *Drosophila melanogaster* genome. We determined the genome-wide occupancy of the mesodermal differentiation factor Twist in the early *Drosophila* embryo using chromatin immunoprecipitation coupled to deep sequencing (ChIP–seq). *In vivo* binding of Twist correlated tightly with the limits of known enhancers. We also tested 31 new candidate CRMs, 21 of which supported expression in a classic dorsoventral pattern or a subregion.

Twist belongs to a large bHLH family of DNA-binding factors that recognize a core DNA consensus, CANNTG, called an E-box. Our analysis showed that *in vivo* and *in vitro* binding preferences of Twist differ: we identified high enrichment of CABVTG motif located within 50 bp of the ChIP summit

and, of these, CACATG was most prevalent. Our mutagenesis experiments with the *rho* CRM further demonstrated that E-boxes CACATG and CATATG, located five nucleotides apart, were not equivalent. Interactions between Twist and other transcription factors might exist; in our data we found only Snail exhibited significant motif co-enrichment in Twist ChIP regions. This finding was neither surprising nor definitive because Snail can bind to a sequence similar to that of Twist.

Among the novel *cis*-regulatory modules identified in Twist ChIP-seq experiments, one was sharing similar spatiotemporal profiles to previously characterized CRMs; including one that shares close similarity with the *sna* expression pattern. *sna* encodes a transcription factor containing Zn-finger DNA-binding domain that predominantly functions to repress expression of a number of genes from ventral regions of the embryo. We showed by *lacZ* reporter assays that the newly identified distal CRM that is located ~7 kb upstream of *sna* gene supported expression that was refined, sharp and similar to the endogenous *sna* expression pattern. To provide insight into *cis*-regulatory mechanisms regulating *sna* expression, we used recombineering methods and created a 25 kb P[acman] rescue construct encompassing the *sna* gene, as well as the flanking DNA sequences. Deletion analysis of the *sna* BAC constructs showed that the distal CRM was necessary to support gastrulation as well as viability of *sna* mutants, whereas the previously identified proximal CRM was dispensable for viability except at high temperature.

Our data also supports the view that although proximal and distal *sna* CRMs probably share many transcription factors, the distal CRM alone was responsive to the repressor Huckebein and the unknown laterally acting repressor, whereas the proximal CRM alone responded to an anterior activator. Overall, our data shows that expression of the *Drosophila sna* gene in the embryo is established through integrated activity of multiple CRMs that function concurrently and, in part, through non-additive interactions.

Taken together, the work presented in this thesis provides evidence that helps illuminate the complex language of chromatin regulation.

Histonen zijn onderhevig aan post-translationele modificaties die in verband worden gebracht met een groot aantal levensprocessen vanwege hun correlatie met gen expressie. Men heeft lang gedacht dat post-translationele modificaties die voorkomen op de N-terminale aminozuur volgordes van histonen de binding van niet-histon eiwitten en eiwitcomplexen aan chromatine moduleren. Sommige modificaties echter, zoals acetylering en fosforylering , kunnen de elektrische lading van histonstaarten veranderen en dus potentieel het chromatine door electrostatische mechanismes beïnvloeden. In dit proefschrift bewijzen we dat de lysine 56 van histon H3 geacetyleerd wordt (H3-K56Ac). Onze resultaten wijzen op een rol voor H3-K56Ac in de reparatie van DNA-schade omdat gist cellen met een K56R mutatie een gevoeligheid tegenover een reeks van DNA schadelijke stoffen tonen. Om de acetylase te identificeren die verantwoordelijk voor de acetylering van H3K56 is, hebben we een specifiek antilichaam gecreëerd dat H3-K56Ac herkent. Een screening van gist stammen met deletie mutaties in 16 bekende eiwit acetylases laat zien dat het niveau van acetylering van H3-K56 niet beinvloed wordt door het verlies van één van de gescreende acetylases. Dit suggereert ofwel dat een tot nog toe onbekende histon acetyl transferase (HAT) bestaat ofwel dat verschillende HATs lysine 56 van histon H3 kunnen acetyleren. Door middel van kleuringen met antilichamen hebben we laten zien dat het niveau van H3-K56 acetylering toeneemt tijdens de S-fase van de cel deling terwijl er veel minder geacetyleerd H3-K56 gevonden wordt in cellen in de G1 fase. Dit wijst erop dat deze modificatie tijdens het plaatsen van de histone op het chromatine plaatsvindt.

Gen expressie wordt op transcriptie niveau gereguleerd door zogenaamde *cis*–regulatory modules (CRM) die door specifieke transcriptie factoren (TF) gebonden worden. Recente studies hebben het *in vivo* bindingsgedrag van een aantal transcriptie factoren op het *Drosophila melanogaster* genoom geïdentificeerd. Wij hebben de binding van de mesodermale differentiatie factor Twist op het gehele genoom van vroege *Drosophila* embryos bepaald met behulp van een combinatie van chromatine immunoprecipitatie en sequencering (ChIP-seq). Dit toonde dat in vivo binding van Twist sterke correlatie vertoont met de grens

van bekende enhancers. Ook hebben we 31 kandidaat-CRMs getest, waarvan 21 expressie in het klassiek dorso-ventrale patroon of een subregio ervan stimuleren.

Twist behoort tot een grote bHLH transcriptie factor familie van DNA-bindende eiwitten die een kern sequentie inhoudende CANNTG (de zogenaamde E-box) herkennen. De zink vinger transcriptie factor Snail herkent ook zulkgelijke motieven. Onze analyse toont dat de voorkeur van Twist voor bepaalde sequenties *in vivo* en *in vitro* verschilt: we vinden een sterke verrijking van het CABCTG motief binnen een afstand van 50 bp rondom de ChIP peak summits en van dit motief was de sequentie CACATG het meest voorkomend. Onze mutagenese experimenten met de E-box CRM van het *Drosophila* gen *rho* hebben getoond dat de E-box motieven CACATG en CATATG die vijf nucleotides uit elkaar liggen functioneel niet equivalent zijn. Mogelijk bestaan er dus interacties tussen Twist en andere bHLH of niet bHLH transcriptie factoren, hoewel onze ChIP resultaten alleen een significante verrijking van Snail in regios waar Twist verrijkt is laten zien. Dit is niet verrassend noch doorslaggevend aangezien Snail een soortgelijke sequentie als Twist herkent.

Een van de CRM die nieuw geïdentificeerd zijn met behulp van onze Twist ChIP-seq proeven toont vergelijkbare spatio-temporale patronen als eerder gekarakteriseerde CRMs, waaronder er één veel gelijkenis vertoont met het *sna* expressiepatroon. Het *sna* gen produceert een Zn-finger DNA bindend domain (omvattende een transcriptie factor) dat vooral een rol heeft in de onderdrukking van de transcriptie van een reeks van genen in de ventrale regio van het embryo. Door middel van *lacZ* reporter assays hebben we aangetoond dat het nieuw geïdentificeerde distale CRM motief, dat ca. 7 kb distaal van het *sna* gen gelocaliseerd is, de gen expressie bevordert, op een manier die vergelijkbaar is met endogene *sna* expressie en net zo gedefinieerd is. Om het *cis*-regulatoire mechanisme dat de *sna* expressie reguleert beter te begrijpen gebruikten we recombinatie methodes en creëerden we een 25 kb P[acman] rescue construct dat het *sna* gen omvat alsmede de naburige DNA sequenties . Deleties in dit *sna* BAC construct tonen dat distale CRM sequenties nodig zijn voor gastrulatie en de levensvatbaarheid van sna mutanten, terwijl de eerder geïdentificeerde proximale

CRM sequentie niet noodzakelijk was voor hun levensvatbaarheid, behalve bij hoge temperaturen.

Onze data steunen de theorie dat ofschoon distale en proximale *sna* CRM sequenties door een aantal van dezelfde transcriptie factoren gebonden worden, alleen het distale CRM motief door de repressie factor Huckebein en een onbekende lateraal functionerende repressie factor gebonden wordt, terwijl alleen de proximale CRM door een anterieure activator gebonden wordt. Globaal laten onze data zien dat expressie van het *Drosophila sna* gen in het embryo gerealiseerd wordt door de geintegreerde activiteit van een aantal CRMs die gelijktijdig actief zijn, maar ook gedeeltelijk door niet-cumulatieve interacties.

De resultaten verzameld in dit proefschrift leveren nieuwe inzichten die helpen de complexe taal van chromatine regulatie beter te begrijpen.

First and foremost I wish to offer my sincerest gratitude to my supervisor Colin. You were always supportive and enthusiastic since the days I began working in the lab as a summer student. I would like to thank you for allowing me the freedom to learn and grow as a scientist. And during the most difficult times when finishing this thesis, you provided the moral support that I needed to move on. I would like to thank to my promoter, Henk for the constant guidance (scientific or not) and support you provided over the years. I am most grateful for your knowledge and patience that helped me become where I am today.

I would like to thank Marion and Gert Jan who provided invaluable advice and insight throughout my graduate work.

Maria, Josephine and Anita (a.k.a Golden Girls), you were most welcoming, always supportive and helpful. It was a privilege to know you ladies and you have been missed.

Joke, I was lucky to have you as a baymate, thank you. Anne Marie, your enthusiasm to teach me the Dutch language and culture was always appreciated. I cannot thank you enough for the excellent technical support that you provided for the lab.

Salva, you were an immense source of knowledge and inspiration. Times we squeezed for a quick chess game or a table tennis match to cheer up were most motivating, thank you. I thank to Jo for her continuing enthusiasm and encouragement. Cinzia, it was a joy having you next door; your work discipline and knowledge about science was very motivating. You even tolerated my little zoo that I created on my desk, even at times it got a little too noisy or distracting.

Zainab, you provided constant moral support and much appreciated insight about life in general. I still miss our pizza & movie nights, Germany trips and night outs. It was always hilarious to find out we had rented the movie several times already that we didn't remember. Rike, I am glad that we became friends. You were most encouraging and persistent that I should finish my degree even at times it seemed impossible, thank you.

Siebe, your assistance in maintaining general equipment was much appreciated, thank you. Celine, I don't remember when we decided to get certified

in scuba diving, but I enjoyed our trips very much.

Suna, your support has always been appreciated; I am lucky to have a lifelong friend like you. Ajda, Berke and Mert, I couldn't have done it without your support, thank you.

Lastly, I would like to thank to my family for their constant encouragement and support in all my pursuits. Thank you.

Anil Ozdemir was born on October18th, 1978 in Zonguldak (Turkey). He studied Molecular Biology and Genetics at Bogazici University (Istanbul), graduating in 2001. He then carried out his PhD at Radboud University Nijmegen, working on regulation of higher order of chromatin structure and histone modifications in yeast. During his postdoc at the California Institute of Technology (Pasadena, California), he began working on mechanisms regulating gene expression during early *Drosophila* development. Since 2006, he lives in Los Angeles, California.

Ozdemir A, Ma L, White KP, and Stathopoulos A. Su(H)-Mediated Repression Positions Gene Boundaries along the Dorsal-Ventral Axis of Drosophila Embryos. *Dev Cell* 31: 100-13 (2014).

Jin H, Stojnic R, Adryan B, Ozdemir A, Stathopoulos A, and Frash M. Genome-wide screens for in vivo Tinman binding sites identify cardiac enhancers with diverse functional architectures. *PLoS Genetics* 9: e 1003195 (2013).

Ozdemir A, and Stathopoulos A. Exciting times: bountiful data to facilitate studies of cis-regulatory control. *Nature Methods* 8:1016-7 (2011).

Dunipace L, Ozdemir A, and Stathopoulos A. Complex interactions between cis-regulatory modules in native conformation are critical for Drosophila snail expression. *Development* 138:4075-84 (2011).

Ozdemir A, Fisher-Aylor KI, Pepke S, Samanta M, Dunipace L, McCue K, Zeng L, Ogawa N, Wold BJ, and Stathopoulos A. High resolution mapping of Twist to DNA in Drosophila embryos: Efficient functional analysis and evolutionary conservation. *Genome Research* 21:566-77 (2011).

Ozdemir A, Masumoto H, Fitzjohn P, Verreault A, and Logie C. Histone H3 lysine 56 acetylation: a new twist in the chromosome cycle. *Cell Cycle* 5:2602-8 (2006).

Ozdemir A, Spicuglia S, Lasonder E, Vermeulen M, Stunnenberg HG, and Logie C. Characterization of lysine 56 of histone H3 as an acetylation site in S.cerevisiae. *Journal of Biological Chemistry* 280:25949-52 (2005).

Cinaroglu A, Ozmen Y, Ozdemir A, Ozcan F, Ergorul C, Cayirlioglu P, Hicks D, and Bugra K. Expression and possible function of fibroblast growth factor 9 (FGF9) and its cognate receptors FGFR2 and FGFR3 in postnatal and adult retina. *Journal of Neuroscience Research* 79:329-39 (2005).